DOCUMENT RESUME

ED 163 085 TM 008 225

AUTHOR Conrad, Linda, Ed.; And Others
TITLE Graduate Record Examinations. Technical Manual.
INSTITUTION Educational Testing Service, Princeton, N.J.
PUB DATE Sep 77
NOTE 112p.; Small print marginally legible
AVAILABLE FROM Graduate Record Examinations, Educational Testing Service, Princeton, New Jersey 08541 ($6.00)

EDRS PRICE MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS Aptitude Tests; *College Entrance Examinations; Content Analysis; Graduate Study; Higher Education; *Literature Reviews; Manuals; Scores; *Statistical Data; Test Construction; *Test Interpretation; Test Reliability; *Test Validity
IDENTIFIERS *Graduate Record Examinations; *Test Manuals

ABSTRACT

This manual supplements previous guides to the use of the Graduate Record Examinations (GRE). It provides sufficient detailed information about the GREs to permit measurement specialists and institutional researchers, as well as faculty members and administrators, to understand the development of the tests and to evaluate their usefulness. Chapters include: (1) A Brief Historical Review of the Graduate Record Examinations; (2) Purposes and General Characteristics of the Aptitude Test and Advanced Tests; (3) Development of the Aptitude Test; (4) Development of the Advanced Tests; (5) Statistical Methods and Analyses of the Graduate Record Examinations; and (6) Validity of the Graduate Record Examinations. Appendixes include information unique to the Aptitude Test and to each of the 20 Advanced Tests including item types, test specifications, norms, and a variety of summary statistics. (ROF)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

GRE

GRADUATE RECORD EXAMINATIONS

TECHNICAL MANUAL

edited by
Linda Conrad
Donald Trismen
Ruth Miller

EDUCATIONAL TESTING SERVICE, PRINCETON, NJ

ED163085

TM008 225

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY
Educational Testing Service
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

GRADUATE RECORD EXAMINATIONS TECHNICAL MANUAL

edited by

Linda Conrad, Donald Trismen, and Ruth Miller

PRIMARY AUTHORS

Robert Altman
Linda Conrad
Ronald Flaugher
Raymond Thompson
Madeline Wallmark
Warren Willingham

EDUCATIONAL TESTING SERVICE • PRINCETON, NEW JERSEY

Reviews or other assistance at various stages in preparation of the text were provided by: Richard Burns, Eleanor Colclough, Beth Drake, Robin Durso, Barbara Esser, Susan Jackson, Miles McPeek, Luis Nieves, Donald Powers, Gary Saretzky, Janis Somerville, Elizabeth Stewart, Stanford von Mayrhauser, Cheryl Wild, Irene Williams, and John Winterbottom. Final reviews were made by Thomas Donlon and William B. Schrader of Educational Testing Service, by Gerald V. Lannholm, formerly of ETS and one of the first directors of the Graduate Record Examinations Program, by Gene Glass of the University of Colorado, and by Nancy Cole of the University of Pittsburgh.

Copies of this publication are available for $6 per copy from Graduate Record Examinations Program, Educational Testing Service, Princeton, NJ 08541. Please enclose payment with your order.

The Graduate Record Examinations Board and Educational Testing Service are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

Copyright © 1977 by Educational Testing Service. All rights reserved.

Library of Congress Catalogue Card Number: 78-66796

4

FOREWORD

The Graduate Record Examinations *Technical Manual* is intended to supplement the *Guide to the Use of the Graduate Record Examinations*, which itself has served as a technical manual for GRE users for a number of years. The *Guide* contains information essential to test users concerning the GRE score scale, limits to the accuracy of scores, and test validity; and it sets forth guidelines to appropriate use of scores, discusses special score interpretation problems, and presents tables of interpretive data based on the total GRE population as well as on subgroups defined by educational level and by major field. Also included in the *Guide* is background material describing the GRE Board, the policy-making body of the testing program, and the various services and publications of the GRE Program.

The purpose of the *Technical Manual* is to provide sufficient detailed information about the GRE to permit measurement specialists and institutional researchers, as well as faculty members and administrators, to understand fully the development of the tests and to evaluate their usefulness.

In 1977–78, a year when departments and institutions are evaluating the new analytical ability measure in the Aptitude Test for its usefulness as an indicator of yet another aspect of developed ability, publication of the *Technical Manual* is considered particularly important. It describes some of the extensive research that led to the introduction of the first major change in the Aptitude Test since the 1940s, when the GRE verbal and quantitative ability measures were introduced.

The *Manual* has been written, insofar as possible, in nontechnical language so that it may be read by the average test user as well as by specialists in measurement. Every effort has been made to include all details necessary for careful evaluation of the Graduate Record Examinations. However, GRE research still in progress and the findings of other testing programs have been generally considered outside the purview of the *Manual*. Likewise, an exhaustive history of the GRE Program has not been attempted; rather a brief historical review sets the stage for a discussion that focuses on the current tests from a historical perspective. The descriptions provided in this edition of the *Manual*, while intended to be definitive, cannot be considered final. As the *Manual* itself reflects, the Graduate Record Examinations have shown a continuous history of change, growth, and adaptation. As data become available on the new analytical ability measure and as the results of recent research can be reported, supplements to the *Manual* will be prepared and distributed.

Richard Armitage, *Chairman*, GRE Board
Lyle Jones, *Chairman*, GRE Board Research Committee

September 1977

CONTENTS

Foreword ... iii

Chapters

1. **A Brief Historical Review of the Graduate Record Examinations** ... 1

2. **Purposes and General Characteristics of the Aptitude Test and Advanced Tests** ... 4
   - Multiple-Choice Format ... 4
   - Test Instructions ... 4
   - Formula Scoring ... 4
   - Test Development Staff ... 5
     - Test Development Procedures ... 5
     - Test Assembly ... 6
   - Quality Control ... 7
   - Testing Standards ... 7
   - References ... 8

3. **Development of the Aptitude Test** ... 9
   - Evolution of the Aptitude Test ... 9
   - General Format ... 10
   - Content Characteristics ... 10
     - Verbal Ability Measure ... 10
     - Content Specifications for the Verbal Ability Measure ... 13
     - Quantitative Ability Measure ... 14
     - Content Specifications for the Quantitative Ability Measure ... 16
     - Analytical Ability Measure ... 16
     - Content Specifications for the Analytical Ability Measure ... 18
   - Statistical Characteristics ... 18
     - Statistical Specifications ... 20
   - Relationship of Statistical Analysis and Research to Test Specifications ... 20
     - Standard Activities ... 20
     - Special Activities ... 20
     - Research Related to Restructuring the Aptitude Test ... 21
   - References ... 23

4. **Development of the Advanced Tests** ... 24
   - Uses ... 24
   - Format ... 25
   - Committees of Examiners ... 26
   - Content Specifications ... 27
   - Statistical Specifications and Characteristics ... 28
   - Information Unique to Each Advanced Test ... 31
   - Reference ... 31

5. **Statistical Methods and Analyses of the Graduate Record Examinations** ... 32
   - Development of the GRE Scaled-Score System ... 32
     - Scaling of the Analytical Ability Measure ... 33
     - Score Equating and Related Concerns ... 34
     - Subscore Scaling ... 36
     - Stability of the Scale ... 36
     - Stability of the Scaled-Score System ... 37
     - Rescaling Study of 1967–68 ... 38
     - Reliability and Error of Measurement ... 39
   - Item and Test Analysis ... 39
     - Item Analysis ... 39
     - Test Analysis ... 41

- Descriptive Statistics 45
  - Basic Normative Data 46
  - Descriptive Statistics for the Aptitude Test 48
  - Other Factors Interacting with Aptitude Test Performance 49
- References 51

6. Validity of the Graduate Record Examinations 52
- Content Validity 52
- Construct Validity 53
- Criterion-Related Validity 54
  - Predictive Validity 54
  - Other Evidence of Criterion-Related Validity 58
- Population Validity 58
- Validity Studies Summarized in Discussion of Predictive Validity 60
- References 63

Appendixes

I. Four Types of Questions Studied but Not Selected for Use in the Analytical Ability Measure of the GRE Aptitude Test 64
- Letter Sets 64
- Logical Reasoning 64
- Evaluation of Evidence 66
- Deductive Reasoning 66

II. Information Unique to Each Advanced Test 68
- Biology 68
- Chemistry 70
- Computer Science 72
- Economics 73
- Education 74
- Engineering 76
- French 78
- Geography 79
- Geology 80
- German 82
- History 83
- Literature in English 85
- Mathematics 87
- Music 89
- Philosophy 90
- Physics 92
- Political Science 94
- Psychology 95
- Sociology 100
- Spanish 101
- References 102

Index 103

FIGURES AND TABLES

Figures

1: Level of Performance of NSF First-Year Engineering Applicants 37
2: Level of Performance of NSF First-Year Mathematics Applicants 38
3: Item Analysis Sample 40
4: Criterion Score Conversion 40

5: Relationship between P+ and Δ when M_{total} = 13.0 41
6: Usefulness of GRE Advanced Test Scores for Predicting Ph.D. Attainment in Three Fields (Creager, 1965) 57
7: Equated Difficulty of Four Types of Reading Comprehension Items in the GRE Aptitude Test for Black Females and a Representative Reference Sample 60

Tables

1: Specifications for Discrete Verbal Questions 13
2: Specifications for Reading Comprehension Passages 14
3: Specifications for Reading Comprehension Questions 14
4: Quantitative Specifications for the GRE Aptitude Test 16
5: Statistical Characteristics of Five Recent Prior Forms of the GRE Aptitude Test 19
5A: Statistical Characteristics of the First Two Restructured Forms of the GRE Aptitude Test 19
6: Statistical Specifications for the GRE Aptitude Test 20
7: A Comparison of Various Experimental Question Types 23
8: Policies of Graduate School Departments Listed in the *Graduate Programs and Admissions Manual* on Use of the GRE Advanced Tests for the 20 Fields of Study Whose Names Match or Closely Match Those of the Tests 25
9: Number of Questions in GRE Advanced Tests and Average Testing Time per Question 26
10: Statistical Characteristics of the Advanced Test Total Scores 28
11: Statistical Characteristics of the Advanced Test Subscores 29
12: Correlations among Scores on the Advanced Tests for which Subscores Are Reported 30
13: Correlations of Advanced Test Scores with Aptitude Test Scores, 1967–68 31
14: Advanced Tests Available 31
15: Scaled-Score Means and Standard Deviations of the 1952 Standardization Group 33
16: Scaled-Score Means and Standard Deviations of the 1967–68 Rescaling Samples 38
17: Total Score Distributions 42
18: Subscore Distributions 43
19A: Scoring Formulas and Reliability Coefficients 43
19B: Intercorrelations 43
19C: Speededness of Test 43
20: Score Distributions 44
21: Frequency Distributions of Original Deltas and Biserial Correlations, by Score 45
22: Item Distribution Sheet 45
23: Summary Statistics for Total Score 46
24: Summary Statistics for Subscores 47
25: Frequency Distributions for All 1975–76 Examinees Who Intended to Major in Microbiology 47
26: 1970–71 Examinee Volume for the Advanced Tests, by Educational Level 48
27: Aptitude Test Performance of Seniors and Nonenrolled College Graduates Classified by Undergraduate Major Field 48
28: Aptitude Test Performance of Seniors and Nonenrolled College Graduates Classified by Intended Graduate Major Field 50
29: Summary Statistics for the Aptitude Test Performance of Seniors and Nonenrolled College Graduates Tested in October 1977, Classified by Undergraduate Major Field 51
30: Aptitude Test Performance of Seniors and Nonenrolled College Graduates, Classified by Graduate Degree Objective 51
31: Aptitude Test Performance of Seniors and Nonenrolled College Graduates, Classified by Citizenship and by Primary Language 51
32: Median Validity Coefficients for Various Predictors and Criteria of Success in Graduate School 56
33: Median Validity Coefficients for Five Predictors of Graduate Success in Nine Fields 56
34: Correlations of Verbal and Quantitative Ability Scores with Self-Reported Undergraduate Grades, and Related Scaled Score Means and Standard Deviations 58
35: Correlations of Verbal and Quantitative Ability Scores with Self-Reported Undergraduate Grades, and Related Scaled Score Means and Standard Deviations, December 1975 58

Chapter 1

A BRIEF HISTORICAL REVIEW OF THE GRADUATE RECORD EXAMINATIONS

The Graduate Record Examinations, known as the Cooperative Graduate Testing Program until 1940, were an outgrowth of a project funded by the Carnegie Foundation for the Advancement of Teaching in the early 1930s to study the outcomes of college education. This project, the "Pennsylvania Study," was the first large-scale attempt to measure academic achievement in higher education by the use of objective multiple-choice tests.

Anticipating a large increase in numbers of applicants for graduate study as the Depression came to an end, the Carnegie Foundation and Columbia, Harvard, Princeton, and Yale Universities continued the work with financial support from the Carnegie Corporation of New York. Faculty committees drawn from the four universities developed tests intended to measure students' intellectual growth and development both through study of the liberal arts and through mastery of specialized fields.

The original test battery consisted of eight "profile" tests in mathematics, physics, chemistry, biology, social studies, literature, fine arts, and a "verbal factor." These tests were administered for the first time in October 1937 to first-year graduate students at Columbia, Harvard, Princeton, and Yale. Since the profile tests were not completely appropriate for measuring one's state of learning in a discipline or major field of study, work was begun to develop 16 "advanced" tests. These were first administered in the fall of 1939.

Interest in the tests spread rapidly. In the early years of the program, validity studies were carried out at the four initial participating universities and also at Indiana and Vanderbilt Universities, the State University of Iowa, and the Universities of Michigan, Pittsburgh, and Wisconsin. By 1940, the tests had been administered to more than 27,000 students in 14 graduate and 26 undergraduate institutions, and results were promising enough to cause widespread consideration to be given to the use of GRE scores as part of the credentials to be presented for admittance to graduate school. In 1942, the Carnegie Corporation said in its annual report that "the examination scores alone are approximately as useful as transcript records taken alone, and the two combined in a manner which uses the test results as a supplement to other evidence of students' qualifications yield a better basis for classifying students than either one used alone" (quoted by Howard J. Savage in *Fruit of an Impulse*, p. 291).

The Testing Modes

Prior to 1942, the Graduate Record Examinations were given solely through "cooperating" institutions—that is, in the so-called "institutional mode." In 1942, however, the increasing use of the examinations as part of the process of admission to graduate study led to the establishment of the first test centers at which students not enrolled in the testing institution could take the tests. The gradual shift toward use in admissions was reflected in the number of undergraduate and graduate students tested: in 1938–39, 1,131 (28 percent) of the 3,869 students tested were undergraduates, while by 1941–42, undergraduates accounted for 5,312 (67 percent) of the 7,936 students taking the GRE. The Independent Student Testing Program was therefore initiated in 1942–43, and in the first year 135 students were tested via the "individual mode" at 35 testing locations.

After the Second World War, as the number of students returning to academic study increased, so did the number of students taking the Graduate Record Examinations. In 1944–45, 6,446 students took the GRE; by 1948–49, the annual number had grown to 51,231

During the same period, the emphasis of the Institutional Testing Program shifted. Initially, it was the mechanism through which graduate institutions tested their own enrolled first-year graduate students, but over the years it was increasingly used by institutions to assess the educational accomplishments of their undergraduate students. To accomodate this particular need of undergraduate schools, the Tests of General Education were introduced in 1946. According to the February 1947 *Bulletin* of the Graduate Record Examinations, the new instruments were designed "to measure as directly as possible the attainment of important objectives of general education at the college level" (p. 8). The Profile Tests continued to be offered through both the Independent Student Testing Program and the Institutional Testing Program, but their use was to be "restricted to graduate and professional students and to applicants to such schools," according to the *Bulletin* (p. 7); and "undergraduate colleges administering the Graduate Record Examinations for purposes of general guidance and appraisal are required to administer the Tests of General Education rather than the Profile Tests" (p. 8).

Content of the Testing Program

In 1949, the GRE Aptitude Test was introduced as a regular part of the Graduate Record Examinations Program, leading to modifications in both the Profile Tests and Tests of General Education as their emphasis on general verbal and quantitative abilities was reduced. The Aptitude Test, first administered as the Graduate Aptitude Test in a 1946 experiment, generated two scores: a verbal ability score and a quantitative ability score. With its introduction, the last basic piece of the Graduate Record Examinations Program as it is known today was in place.

In January 1948, the Graduate Record Examinations became the responsibility of the newly established Educational Testing Service. Almost immediately, liaison was established with a newly created Committee on Testing of the Association of Graduate Schools (AGS) in the Association of American Universities, which worked with the GRE Program office to review the tests and services offered. In 1951, the name of the Independent Student Testing Program was changed to the National Program for Graduate School Selection, and changes in the test offerings continued as the needs of both the National Program and Institutional Program were continually reevaluated.

With the growth in the utility and use of the new Aptitude Test, the Profile Tests were discontinued in 1953 in the National Program and in 1954 in the Institutional Testing Program. In that same year, the Institutional Testing Program also discontinued the Tests of General Education, replacing them with the Area Tests, a comprehensive appraisal of college students' orientation in three principal areas of human culture: social science, humanities, and natural science. Over the course of several years, Advanced Tests in education, engineering, and Spanish were introduced, and Advanced Tests in agriculture, fine arts, German, and home economics were discontinued.

By 1964, the GRE Program included the Aptitude Test and 18 Advanced Tests in biology, business, chemistry, economics, education, engineering, French, geology, government, history, literature, mathematics, philosophy, physical education, physics, psychology, sociology, and Spanish. In 1965, new Advanced Tests in music and speech were introduced; in subsequent years Advanced Tests in geography (1966), anthropology (1968), and German (1970), were introduced, and in 1970 the tests in business and physical education were discontinued. After reconsideration, the new tests in speech and anthropology were also discontinued in 1970 and 1971, respectively.

By 1972, the basic GRE test offerings had evolved into 19 Advanced Tests and the Aptitude Test with verbal and quantitative sections. Since that date, approximately 300,000 people have taken the Aptitude Test each year, with mean scores based on all test-takers ranging between 497 and 492 over the five-year period since 1972. A complete review of the existing Advanced Tests was carried out between 1970 and 1972, resulting in the availability of subscores as well as total scores for nine of the tests. Then, beginning in 1974, a series of studies relating to the possible restructuring of the Aptitude Test was undertaken; the result, in October 1977, was a new Aptitude Test including the basic verbal and quantitative ability measures and a measure of analytical ability as well. Thus, as of 1977, the program's basic test offerings include the Aptitude Test with verbal, quantitative, and analytical sections and 20 Advanced Tests, the twentieth, computer science, having been added in 1976.

The GRE Board

By the mid-1960s, the growing importance of the Graduate Record Examinations created a need for greater participation by the graduate school community in setting policies. In 1965, ETS, the AGS Committee on Testing, and the Committee on Testing of the Council of Graduate Schools (CGS) in the United States jointly prepared a Proposal for the Establishment of a Graduate Record Examinations Board. The proposal began with the following statement: "The use of the GRE has increased significantly and, although ETS has always sought the advice of the graduate schools and their faculties in the development and administration of the program—most notably through the AGS Committee on Testing—the increasing use of the GRE, and the likelihood that the increasing trend towards graduate study will accelerate that use, makes it appropriate that a closer relationship between the graduate schools and the GRE be considered" (p. 1).

As a result of that proposal and actions taken by the Executive Committees of both AGS and CGS, the Graduate Record Examinations Board was created effective January 1, 1966. Consisting of four members appointed by AGS, four appointed by CGS, and eight appointed at large by the board itself, the new board soon signed a compact with ETS that outlined the board's responsibilities and ETS's agreement "to vest in the board authority over the general policies of the GRE" (p. 1).

Since 1966 all major decisions concerning the offerings and administration of the GRE Program have been made by the GRE Board, drawing on professional personnel at ETS as board staff. One of the board's first major decisions was to limit its concerns and policy control to the National Program for Graduate School Selection, and this decision was implemented in October 1969. (The Institutional Program was continued by Educational Testing Service as the Undergraduate Program.) Effective in October 1969 also, the GRE Board instituted the Local Administration Service to enable graduate schools to administer the GRE to their own enrolled graduate students for purposes of evaluating students or programs, selecting students for more advanced programs, and other nonadmission purposes. (The Local Administration Service will be discontinued after June 1979 because of declining graduate-school interest.) The Special Administration Service, which enables students to take the GRE on dates other than those of National Administrations, had begun in New York City in the 1940s. Additional Special Administration test centers in other large cities were opened in subsequent years, to a total five by 1967 and eight by 1973.

As provided for by its bylaws, the board soon developed a committee structure and, reflecting a priority that has continued up to the present time, created as its first standing committee, a Committee on Research and Test Development. Board-sponsored research projects addressed a wide range of issues and questions relating to the transition from undergraduate to graduate study and graduate study itself, as well as matters more directly related to the GRE Program of tests and services. Early studies included investigation of alternate methods of equating GRE tests (1969) and of the use of Bayesian statistics to facilitate validity studies (1970), as well as surveys of existing graduate admissions and fellowship selection policies (1970) and of programs available for disadvantaged students (1973). More recently, board projects have included studies of male and female doctoral students, the identification of dimensions of quality in doctoral programs, and several projects relating to possible modification of the GRE Aptitude Test.

Other Activities of the GRE Board

In 1973 the GRE Board entered into an agreement with the College Entrance Examination Board and ETS to create a new policy group for the Test of English as a Foreign Language (TOEFL) to which the GRE Board appoints three representatives. In similar fashion, the GRE Board participates in the governance of the Graduate and Professional School Financial Aid Service (GAPSFAS) by appointing three representatives to its Council. In 1976, as an outgrowth of a special study committee jointly created by ETS and the GRE Board, the Undergraduate Assessment Program (UAP) Council was created to assume policy direction for that program, the revised and redesigned descendant of the former GRE Institutional Testing Program that had been administered by ETS as the Undergraduate Program since 1969.

Between 1937 and 1967, activities relating to the Graduate Record Examinations focused almost exclusively on the tests and the arrangements for their administration; since 1967, the GRE Board has broadened the services and activities of the GRE Program to include research, guidance publications (*Graduate Programs and Admissions Manual* [*1972 ff*] *and Thinking about Graduate School* [*1973*]), a Minority Graduate Student Locater Service (1973), and numerous special projects, surveys, and conferences concerning issues in graduate education.

As the Graduate Record Examinations begin their fifth decade—and their second under the direction of the GRE Board—they continue to play an important role in the admissions process to

American graduate education; equally important, they now provide the base from which a broad program of related research, publications, and services has evolved.

References

A Compact between the Graduate Record Examinations Board and Educational Testing Service, April 1968.

A Proposal for the Establishment of a Graduate Record Examinations Board. Prepared by Educational Testing Service in consultation with the Committee on Testing, Association of Graduate Schools, and the Committee on Testing, Council of Graduate Schools in the United States, March 1965.

Savage, H. J. *Fruit of an Impulse: Forty-five years of the Carnegie Foundation 1905–1950.* New York: Harcourt, Brace and Co., 1953.

Vaughn, K. W. *The Graduate Record Examination: A statement of policy to cooperating colleges and universities* (The Graduate Record Office Bulletin Number 1). New York: The Graduate Record Office, February 1947.

Chapter 2

PURPOSES AND GENERAL CHARACTERISTICS OF THE APTITUDE TEST AND ADVANCED TESTS

The GRE Aptitude Test and Advanced Tests are provided to aid prospective graduate students and institutions in the application and admissions process. Students take the tests, usually in response to an institutional or departmental requirement, to provide information in addition to undergraduate grades and other indicators of past and potential performance.

The examinations are administered several times a year, both nationally and in foreign countries, under standardized conditions. Scores are usually reported from four to six weeks after each test administration. Detailed information related to the administrations is found in the GRE *Information Bulletin*, a new edition of which is published annually.

To prevent the contents of a given test from becoming common knowledge and to assure that scores are authentic, three primary methods are used: 1) strict regulation of the handling of test materials so that students will not have an opportunity to see the test except during its administration; 2) provision of multiple parallel forms of the test (the number of forms determined by the number of students taking the particular test) to reduce the chance that a student will take the same test twice and make more difficult possible attempts to divulge test content; and 3) administrative procedures to curb impersonation and copying. In addition, the multiple scores of repeaters (students who take the test more than once) are checked for statistically unlikely differences. Such unusual cases are investigated by the ETS Security Office to determine whether an irregularity has taken place.* Reports from students or supervisors of suspicious behavior are also thoroughly checked. Scores that prove to be inauthentic are either not reported or canceled.

The test scores are intended for use along with other indicators of student performance by graduate departments, graduate admission committees, and fellowship sponsors in making admission decisions or awarding fellowships. The *Guide to the Use of the Graduate Record Examinations,* which this manual supplements, sets forth program policies concerning who may receive the scores. The *Guide* also provides all necessary information for properly interpreting and using the scores reported for the Aptitude Test and Advanced Tests.

Multiple-Choice Format

All tests now offered by the GRE Program are in a multiple-choice format. Although they have certain limitations, modern objective tests can present challenging intellectual tasks to examinees as well as measure factual knowledge. A prime advantage of multiple-choice tests over free-response tests is that they permit wider and, hence, more accurate sampling of material in a given period of time. Thus, more measurements of different facets of examinees' thinking are secured per unit of time. A second important advantage is the elimination of one source of measurement error. In a free-response test, measurement errors are associated with the test, the examinee, and the grader. In multiple-choice tests, one of these sources of error, the grader, is eliminated. These two advantages result in higher reliabilities for multiple-choice measures. A third advantage is their practicality; the tests can be scored by machine so that scores can be reported more quickly and more economically than would otherwise be possible.

Test Instructions

The general instructions for taking the tests are intended to suggest a widely applicable test-taking strategy, to alert the student to the multiple-choice format, and to describe the method of scoring, which corrects for random guessing. These general instructions are provided for the students to read at the beginning of the testing sessions. (The complete text of the instructions is in the GRE Sample Aptitude Test made available by the GRE Program.)

The instructions specific to and immediately preceding groups of each type of question are provided within the timed sections and are made as concise and clear as possible. Questions that have answer choices unique to them tend to require very brief instructions. However, fixed-format questions (those for which a fixed set of answer choices applies to all or to a group of questions) tend to require longer instructions. The fixed-format questions, however, are relatively less time-consuming than most of the unique answer-choice questions. In cases where the method of solving a problem or the criteria a student must use to evaluate material must be established, examples are included as part of the instructions.

Formula Scoring

All questions that contribute to a given score have equal weight. To eliminate the potential advantage of random or haphazard guessing, formula scoring is used. The formula used for computing scores on the tests is $R - \frac{W}{K-1}$, where

R is the number of right answers
W is the number of wrong answers, and
K is the number of answer choices per question.

For most of the tests (all except part of the quantitative section of the Aptitude Test, the Advanced German Test, and part of the Advanced Spanish Test) this formula becomes $R - \frac{W}{4}$ since each question has five answer choices.

The rationale for the use of the instructions on guessing and the use of the scoring formula rests on the most likely result of pure guessing. If an examinee makes random or haphazard guesses on each five-choice question in a test, the most likely result is that one-fifth of the questions will be answered correctly and four-fifths of the questions incorrectly, by chance alone. In the most likely out-

---

*Procedures for protecting the student under investigation and for assuring the authenticity of scores are described in *ETS Procedures for Determining the Validity of Questionable Scores.*

come, application of the formula $R - \frac{W}{4}$ yields the intuitively reasonable score of zero.

All other outcomes, from getting all the questions right to getting all the questions wrong, can also be the result of pure guessing. For example, there is a probability of $\left(\frac{1}{5}\right)^n$ that an examinee would get all n questions in a test of five-choice questions right through pure guessing. That is a vanishingly small probability for a test of 100 or more questions; nevertheless, it is possible to get all questions right through haphazard guessing, and application of the correction formula in that case would not alter this highest possible score. For a random guesser taking a 100-question test, with each question counting 1 raw score point, the probability that a corrected score will be greater than 5 points is only 1 chance in 6, greater than 10 only 1 chance in 40, and greater than 15 points only 1 chance in 740.

As soon as the examinee begins to read the questions and bring some knowledge to bear in answering them, one leaves the realm of pure guessing. If the examinee can rule out one of the five answer choices as wrong and guesses the answer from among the four remaining options, the probability is 1 chance in 4 that the examinee will gain 1 point by responding correctly, and the probability is 3 chances in 4 that the examinee will lose 1/4 point by responding incorrectly. Since $\frac{1}{4} \times 1 = \frac{1}{4}$ and $\frac{3}{4} \times \frac{1}{4} = \frac{3}{16}$, the odds clearly favor answering a question in which one or more of the answer choices can be ruled out as wrong. Thus, the general instructions, to be read before beginning the test, advise the examinee: "It is improbable . . . that mere guessing will improve your score significantly; it may even lower your score, and it does take time. If, however, you are not sure of the correct answer but have some knowledge of the question and are able to eliminate one or more of the answer choices as wrong, your chance of getting the right answer is improved, and on the average it will be to your advantage to answer such a question."

Test Development Staff

The professional staff primarily responsible for the content of the GRE Aptitude Test and Advanced Tests generally have advanced degrees in fields related to the tests they develop; for example, those responsible for the verbal measure in the Aptitude Test tend to have backgrounds in the humanities or in measurement. Those responsible for the mathematics or quantitative portion tend to have advanced degrees in mathematics or a related field. Responsibility for the analytical measure is shared by people with humanities, science, and mathematics backgrounds, some of them with formal training in logic.

Test specialists usually have considerable experience in test development and have training in psychometric principles and techniques as they relate to test construction. Experience in teaching is quite common. Most of the test development staff maintain close contact with experts in their respective fields and have immediate access to test-related research carried out at ETS and conducted outside the organization (the Brigham Library, located at ETS, has a large collection of test-related materials). Persons preparing the reading comprehension materials for the Aptitude Test, for example, regularly receive numerous nontechnical periodicals written at a level appropriate for the students being tested. Test development staff members also have access to the Firestone Library at Princeton University and other nearby educational institution libraries as well as to ERIC, a computer-accessed library of educational research.

In preparing the Advanced Tests in particular fields, test specialists take the initiative in securing new questions, obtaining reviews of new questions and tests, arranging for committee meetings, planning meeting agendas and working schedules, and shepherding a test through assembly, editing, and production. The role of the test specialists in working with the committees of examiners, consisting of experts in the respective fields, partly depends on their background, experience, and personality.

The test specialists represent a wide cross section of educated persons in the United States, including men and women who come from various regions of the country, have different religious and ethnic backgrounds, and have been educated in small as well as large, and public as well as private, institutions. A diversified staff is considered especially important because the tests must be made appropriate for a large, heterogeneous population. Approximately equal numbers of men and women are test specialists for GRE Advanced Tests; minority group members are also represented.

In an attempt to achieve even greater diversity, the test development staff hires writers outside the organization to supplement material generated at ETS. These may include former staff members, recommended advanced graduate students who are close to the academic activities of the population for which the test is developed, and faculty members with a professional interest in testing.

Test Development Procedures

Methods of generating material are unique to each writer, but formal standardized procedures have been developed to guide the generation process, to assure uniformly high quality material, to avoid idiosyncratic questions, and to encourage the development of test material that is widely appealing.

An important part of the generation of test material is the review process. Each question developed, as well as any stimulus material on which questions may be based, must be reviewed by several independent critics. In the review process for the Aptitude Test, the writer must take into consideration a reviewer's comments, revise the question as necessary, submit the revised question for a second review by another individual, again revise, and, if changes are substantial, submit the question for yet a third test specialist's review. In certain cases, questions may be reviewed by an expert outside ETS who can bring a fresh perspective to review of the questions.

Central to the review process for the Advanced Tests is the committee of examiners for each test. After Advanced Test questions have been written, they are reviewed by the ETS test specialists and then prepared in multiple copies for review by the committee members. Each committee member receives a collection of all the questions and forms on which to record reactions. Ordinarily, the committee members are asked to indicate the correct answer for each question, their rating of the importance of the question's subject-matter content, their rating of its technical quality, and any revisions or comments they deem appropriate. Probably the single most important part of the review is their indication of the correct answer. Any disagreements among the committee members regarding the correct answer clearly signal the presence of possibly serious flaws. The ratings of question content and technical quality are also important, and distinguishing between content and technical quality is useful. Questions rated highly on both content and technical quality are clearly the best, and those with high content and low technical quality ratings may well be worth revising to improve their technical quality. However, those with low content ratings are likely candidates for discarding, regardless of their technical quality.

The next step is collation of the independent reviews of the committee members. A copy of the collated reviews is sent to each member. The sources of individual review comments are not identified. Thus, the committee members review the new questions and then have a chance to review the consolidated reviews of their fellow committee members.

The most significant activity at almost every committee meeting is thorough review of questions for a new test edition. Most of the time and energy of the participants is devoted to this activity. Generally, the new questions are taken up one by one. After discussion, each question is either approved (often with substantial revision), discarded, or held for possible revision or use in the future. Then decisions must be made as to which approved questions should be used in the test to provide a balanced coverage of all aspects of the test specifications and to avoid undue overlap.*

It may be useful to think of three facets of question review for the Advanced Tests: review of subject matter, review of technical quality, and review of editorial style. As a rule, the committee members and ETS test specialists are in a position to provide all three facets of the review. It is not uncommon, however, for committee members to focus on subject matter, ETS specialists to concentrate on technical quality, and ETS editors, who review questions at a later stage, to concern themselves mainly with editorial style.

Test Assembly

After the items have been reviewed and revised, they are culled to produce a group of the best questions, consistent with the specifications, for inclusion in a test. In the case of the Aptitude Test, the selected questions are assembled first in pretest form. The questions judged best based on their performance in the pretest become part of a pool of questions from which a final form of the test is assembled. In the case of the Advanced Test, assembly of the final form typically begins in the committee meeting and is completed by the test specialist with committee advice.

The test assembler considers not only the individual question but also the relationship of the question to the entire group of questions in the test being prepared. For example, in preparing the Aptitude Test, the test assembler makes sure that no two questions are actually asking the same thing in a set of reading comprehension questions and avoids in vocabulary questions the frequent reappearance of words already in the test. Test assembly requires ordering the questions, with very easy questions placed first, balancing the questions to meet test specifications for content and statistical qualities, and recording information showing how closely the test matches the specifications.

Test assembly includes attention even to such seemingly minor details as assuring that no preponderance of correct answers is associated with a particular letter (E, for example). The following formula is used to check the balance of correct answer choices: $\frac{N}{n} \pm \sqrt{\frac{N}{n-1}}$ where N is the number of questions in the test and n is the number of options for each question. In a test of 160 five-answer-choice questions, the number of (E) correct answers must be:

$$\frac{160}{5} \pm \sqrt{\frac{160}{4}} = 32 \pm 6 = 26 \text{ to } 38.$$

In a folder with the assembled questions (each on an individual card with information, if available, on its statistical characteristics) is included several kinds of information: 1) the title, assembler, purpose, and schedule of the examination; 2) the content characteristics of the test; 3) the distribution (with means and standard deviations computed) of questions according to estimated or known difficulty and discriminating power (for the entire test and, in the case of the Advanced Tests, for the equating subtest separately); 4) specifications for equating and item analysis; 5) a record of previous sources, uses, and statistical characteristics of each question; and 6) an official key. The official key shows the correct answer for each question. This key can be certified as official and signed by the test assembler only if at least three independent experts have agreed on the correct answer for each question.

When the test has been assembled, it is reviewed by a second test specialist. Then it is reviewed by the test development coordinator. After mutually agreeable resolutions of any points raised in these reviews have been reached, the test goes to a test editor. The test editor's review is likely to result in many suggestions for change, and the test assembler must decide how these suggested changes will be handled. If a suggested change yields an improvement from an editorial viewpoint, without jeopardizing the content integrity, the change is made. Otherwise, new wording is sought that will meet the dual concerns of content integrity and editorial style. After a second careful editorial review by a copyreader, camera-ready planograph copy is prepared by specialists in test typing, drafting, layout, and proofing.

In the case of the Aptitude Test, the camera-ready copy is returned for reviews by the test assembler and by another test specialist. The assembler and planograph reviewer check for any problems that may have been overlooked. All reviewers except the editors, copyreaders, and proofreaders must attempt to answer each question without first checking the answer key. This means that each reviewer is "taking the test" and is uninfluenced by knowledge of what the question writer or test assembler felt the answer should be.

In the case of an Advanced Test, the camera-ready copy must be reviewed again by the committee of examiners. Photocopies of the camera-ready test are sent to each member of the committee of examiners. At this stage the committee members are asked to take the test and to mark the correct answers to the questions. They note any changes they think need to be made to ensure accuracy and eliminate ambiguity. On the basis of these reviews, the test assembler specifies the final changes to be made. Special problems may require consultation with the committee chairman.

After a final review for correspondence between directions and questions, question and page numbering, and overall layout, the planograph is sent to the printer under conditions designed to protect the confidentiality of the test material. Review of a proof copy precedes printing.

---

*For a more detailed discussion of the nature and role of committees of examiners, see Chapter 4.

Quality Control

Test quality and the consistency of quality across test editions are controlled largely through the extensive reviewing process during which a number of independent critics evaluate each question in a test for content, clarity, accuracy, and style. However, two methods requiring statistical analysis of questions are of major importance in assuring that a final test of high quality is produced: pretesting for the Aptitude Test and preliminary item analysis—statistical analysis of individual questions in a final test form before scoring—for the Advanced Tests. A full test analysis that gives detailed information on the test's reliability, score distributions, speededness, and other characteristics is always provided to test development staff and committees after scores have been reported. However, the purpose of a full test analysis is to guide development of future forms of the test and to document the characteristics of a given form. The purpose of pretesting and of preliminary item analysis is to assure that a given test form, by the time it is produced and scored, contains questions that are without serious flaws.

Pretesting requires inclusion of some questions in the test that do not contribute to examinees' scores but are experimentally "scored" for a representative sample of the population to obtain information on the difficulty and usefulness of the questions. All questions contributing to anyone's GRE Aptitude Test scores have been pretested before inclusion in a final form of the test. Pretest data are valuable because they enable test specialists to eliminate poor questions (perhaps revising and repretesting them) and to meet rather precisely the test specifications for difficulty and reliability. Examinees are informed through the GRE *Information Bulletin* and the GRE Sample Aptitude Test that such trial questions are part of the examination and that they will not affect reported scores. After statistical analysis is completed on pretested questions, information on the performance of each question is pasted on a card with the question printed on the front. This assures that the information on the question will be readily available and convenient to use; the card is then used in assembling a final form of a test (assuming that the performance of the question is satisfactory). The analysis on the back of the card provides information on the number of people in the sample, difficulty level of the question (percent answering it correctly), number selecting each answer choice, and mean ability level of those selecting each choice (mean ability level is defined in terms of performance on the appropriate ability measure in the actual test). Another bit of vital information on the analysis card is the r-biserial, that is, the question-test correlation. If students doing well on the test as a whole also do well on the question, the correlation will be relatively high; if not, it will be relatively low. If all the r-biserials are very high, the test may be measuring a construct too limited or too narrowly defined. If the r-biserial of a question is very low, it is not contributing to the reliability of the test as a whole. The low r-biserial suggests that there may be a problem inherent in the question or that students are unfamiliar with the material or concept tested.

The GRE Advanced Tests have typically been constructed without pretesting. Even though pretesting would permit development of forms more nearly parallel in difficulty, some differences among test forms in this respect are acceptable since one edition of a test is statistically equated to other editions. Thus, students taking a more difficult test would not have to answer as high a percentage of questions correctly as those taking an easier edition. For a full year, pretesting was tried for all the Advanced Tests. The process proved to be of little value in improving the overall reliability of the tests, and sectioning the tests to allow for a separately timed pretest section caused administrative problems. Students who otherwise might have finished the tests and left early experienced restlessness when they finished each individual section early.

For other reasons also, pretesting for the Advanced Tests did not prove particularly useful. First, it is easier for a committee of examiners in a field to estimate the difficulty and discriminating power of a question in that field than it is for test specialists to estimate the difficulty of an Aptitude Test question that will be used for a widely heterogeneous group embracing all fields of study. Second, each Advanced Test contains a number of already tried questions (generally 20 percent of the test, sometimes more) included for equating purposes. These questions have all proved to be of appropriate difficulty and high discriminating power.

Preliminary item analysis is an important procedure in controlling the quality of the Advanced Tests. Preliminary item analysis is performed also for the Aptitude Test, as an additional check for such problems as possible misprints, but the previous pretesting step makes the preliminary item analysis less important for the Aptitude Test than for the Advanced Tests. Before a test being administered for the first time is scored, a sample of answer sheets arriving relatively early is experimentally scored and analyzed. A question that reveals poor discriminating power, inordinate difficulty, or a large number of omissions is reviewed again at this point by test specialists and committee members to make certain that the question is not ambiguous and that the answer designated as correct is indeed the only correct answer. If problematical questions are identified that escaped the attention of the committee or test specialists earlier, a decision can be made to eliminate the question from scoring or, possibly, to permit two correct answers. Many Advanced Tests require no change at this stage; others may require action in the case of one or a few questions. Because of the effectiveness of the pretesting process for the Aptitude Test, a change in the scoring instructions is almost never needed. Although the methods of pretesting and preliminary item analysis differ in their importance for the Aptitude Test and the Advanced Tests, these methods are vital to the maintenance of quality in the Graduate Record Examinations and are effective in keeping reliabilities in the very high .80s or above. Pretesting and preliminary item analysis are discussed in more detail in Chapter 5.

Testing Standards

The standards that apply to all GRE tests are summarized below.

1. Tests used to assist in making decisions that are typically irrevocable and have significant impact on students' courses of action should have reliabilities that do not fall below the upper .80s or low .90s. Tests with lower reliabilities can be provided for such purposes as self-evaluation or counseling.
2. All scores used to assist in making significant decisions should be sufficiently distinct to warrant separate reporting (for example, score intercorrelations below .80 when reliabilities are in the .90s).
3. The measures should provide a distribution of scores that approximates the normal curve.
4. The tests should not be highly speeded.
5. The tests should have appropriate content for the constructs they are designed to measure and should be positively correlated with successful performance in graduate school.

6. Sufficient information should be provided to users to permit appropriate interpretation of scores.

References

Educational Testing Service. *ETS procedures for determining the validity of questionable scores*. Princeton, N.J.: Educational Testing Service, 1975.

Educational Testing Service. *GRE 1977–78 Information Bulletin*. National Administrations Edition. Princeton, N.J.: Educational Testing Service, 1977 (published annually).

Educational Testing Service. *GRE Sample Aptitude Test*, second edition. Princeton, N.J.: Educational Testing Service, 1977.

Educational Testing Service. *Guide to the Use of the Graduate Record Examinations, 1977–78*. Princeton, N.J.: Educational Testing Service, 1977.

Chapter 3

DEVELOPMENT OF THE APTITUDE TEST

The GRE Aptitude Test is a standardized test of general academic ability. It includes three measures: verbal ability, quantitative ability, and a newly added analytical ability measure. The Aptitude Test is intended to reflect skills that have developed over a long period of time. Although it assumes exposure to a predominantly English-speaking culture and to the educational practices of the United States, the test is designed to be as appropriate as possible for potential graduate students with diverse backgrounds and interests.

The purpose of the test is to contribute to prediction of a student's performance in graduate school. Not only is it based on constructs that are theoretically related to successful study in a variety of fields, but performance on the test has been demonstrated to be positively correlated with performance in graduate and undergraduate school as measured by various criteria. The Aptitude Test is not intended to measure inherent intellectual capacity or intelligence; nor is it intended to measure personality traits or social worth. Its limited purpose is to tap the ability to reason with words, mathematical concepts, and other abstractions to arrive at a solution to a problem. Such factors as knowledge of words and mathematical concepts and practice in reading and fundamental quantitative operations will, of course, define the limits within which one can reason using these tools.

The rationale for the content of the Aptitude Test originates in the need for highly developed fundamental skills in graduate study of any kind. Three scores rather than a single score are provided for several reasons: 1) a multidimensional definition of academic talent will best serve institutions and students in a variety of fields; 2) the three scores—verbal ability, quantitative ability, and analytical ability—are sufficiently independent to be providing complementary information about students; and 3) studies suggest that each score is related to academic performance in differing degrees depending on the field and may differentially improve prediction of graduate school success.

Evolution of the Aptitude Test

The development and evolution of the Aptitude Test have been determined by perceptions of the needs of the students and institutions making up the graduate community and by high standards of psychometric quality. These perceived needs and established standards—and the fact that they are sometimes in conflict—are reflected in published materials concerning the Aptitude Test in various stages of its evolution. For example, the GRE *General Bulletin*, No. 2, in 1948 noted that "further breakdowns of the verbal ability score are anticipated if analyses of the test results show them to yield satisfactorily reliable and differentiating part scores" (p. 3). Although part scores were perceived as potentially valuable to the graduate community, it was discovered that the various kinds of verbal questions were so highly intercorrelated that part scores could not be defended as psychometrically sound.

Throughout its history, the Aptitude Test has been considered to be relatively independent of passing trends in student interests and teaching methods. Because the primary advantage of a standardized test is its capacity to permit comparison of students by the same standards, only two kinds of change have thus far been introduced into the Aptitude Test: 1) minor change in specifications of content expected to result in a measure more appropriate for the population without compromising parallelism of test forms and comparability of scores over the years; and 2) change that would increase the usefulness of the test without subtracting advantages already offered. In all instances where change has been suggested, results have been analyzed statistically to determine the possible effects.

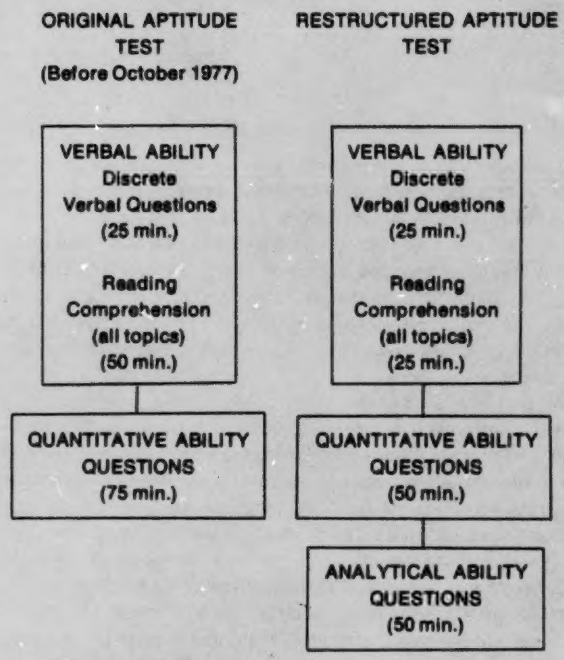
Two recent examples illustrate the kinds of minor change generally accommodated in the Aptitude Test, typical reasons for such minor change, and the results of the change as demonstrated by statistical analysis. Because of current plans in the United States to introduce the metric system, some of the terminology used in the quantitative measure of the Aptitude Test has been altered to reflect changes taking place in the educational system. Some questions previously referring to feet and inches, for example, may now refer to meters and centimeters. However, it is not required that students know the number of inches in a meter or the number of quarts in a liter. The numbers and computation have not changed, but the terms may have. Statistical analysis has shown that changes in the terminology in the questions have not, on the average, affected their difficulty or their usefulness in distinguishing between high and low scorers. Until the new system of units and measurement has become firmly entrenched, the quantitative measure in the Aptitude Test will not require knowledge of its fundamentals.

Recently, the specifications of the verbal ability measure of the Aptitude Test were refined to reflect social change thought to have resulted in a different mix of reading materials in the average student's experience. Diversification of passages in the reading comprehension section had previously been assured by concern for balance among humanities, social studies, and science passages and inclusion of various styles, such as fiction and argumentation. The refinement in specifications added a requirement for one passage relating to minority concerns and one relating to the concerns of women. The purpose of this refinement was to increase the appropriateness of the content for the heterogeneous population and to increase the resemblance of the reading selections in the test to materials available to the typical student. Statistical analysis of the passages with content related to minorities showed that they were not significantly different from other passages in the same general categories—that is, in the humanities, social studies, and science—for the total population.

The feasibility of making major changes in the test (such as the addition of new measures) to increase its usefulness has been investigated periodically. In the early 1950s, a number of potentially useful types of questions were tried out in experimental sections of the GRE Aptitude Test: for example, questions designed to test the "ability to reason logically in terms of abstract figures," as the directions in a 1951 pretest suggested; the ability to interpret data or to judge the sufficiency of data; the ability to "integrate" material in an essentially artificial language with more rules, greater complexity, and less dependence on knowledge of grammar than other such tests; the ability to induce rules in such tasks as completing analogies, completing a series of symbols or concepts, and selecting an incompatible term or symbol in an otherwise logically related series; the ability to judge evidence (these questions resemble a type of question investigated more recently, Evaluation of Evidence—see Appendix I); and even a non-multiple-choice type of question involving categorization of words in lists. These experimental efforts, however, did not lead to expansion of the test.

In the late 1960s and early 1970s, such possible measures as Spatial Visualization, History of Ideas, Writing Skills, High Level Math Usage, and Logical Reasoning were examined with a view of permitting students to select optional measures based on their intended specialization. At that time, the tests judged by a sample of faculty members on the Advanced Test committees of examiners to be most important and potentially useful were the Logical Reasoning Test (assumed to be selected as an option by humanities, social science, and some natural science students) and the High Level Math Usage Test (assumed to be selected as an option by some science students). Despite reliability and promise of validity, scores on the Logical Reasoning Test were too highly correlated with verbal scores to be considered a valuable adjunct to the original Aptitude Test. Although the High Level Math Usage Test was found to be appropriate for its intended use, its introduction depended on shortening the part of the quantitative measure common to all students to only 30 minutes. Such a reduction in time was considered to be unwise because it limited the common measure's diversity. However, the High Level Math Usage Test was incorporated into the Advanced Engineering Test, a measure in which it appeared particularly useful.

In 1974, a new effort was initiated to consider possible improvement of the Aptitude Test by broadening its definition of academic talent. That effort is continuing and includes investigation of methods of testing for scientific creativity and cognitive style. Results of research suggested that the verbal and quantitative portions of the test, as it existed before 1977–78, could be shortened without reducing reliability below a satisfactory level and without affecting the comparability of past scores with scores based on the shortened versions. This research effort also yielded information on a variety of tests of various aspects of reasoning. A subset of those tests was selected to form a new measure of analytical skills. The 1977–78 Aptitude Test differs dramatically from the test that preceded it. However, that difference represents an added value to the test and maintains the importance of the traditional verbal and quantitative measures. The diagram at the right illustrates the difference between the two tests in scores yielded and in basic content.


ORIGINAL APTITUDE TEST (Before October 1977)
RESTRUCTURED APTITUDE TEST
VERBAL ABILITY
Discrete Verbal Questions (25 min.)
Reading Comprehension (all topics) (50 min.)
QUANTITATIVE ABILITY QUESTIONS (75 min.)
VERBAL ABILITY
Discrete Verbal Questions (25 min.)
Reading Comprehension (all topics) (25 min.)
QUANTITATIVE ABILITY QUESTIONS (50 min.)
ANALYTICAL ABILITY QUESTIONS (50 min.)

General Format

The restructured Aptitude Test consists of five separately timed sections, two of which are 50 minutes long and three of which are 25 minutes long. The verbal measure consists of 80 questions; the quantitative measure consists of 55 questions; and the analytical measure consists of 70 questions. Equal amounts of time—50 minutes each—are devoted to the three measures. Twenty-five of the 175 minutes of the students' time is spent answering trial questions.

Although not contributing to the scores of the students who take them, trial questions are considered an integral part of the examination, essential to maintaining the high quality of the test. Unless the trial questions can be given to a sample of the regular GRE population under normal standardized conditions, the statistical data most dependable in making parallel forms of the test cannot be obtained. The trial questions represent research that directly benefits the students who take the tests, and students are, of course, informed in the GRE *Information Bulletin* of the inclusion of trial questions in the test. In addition, the test supervisor, just before the examination is administered, reiterates that "each edition of the Aptitude Test contains a number of questions being tried out or pretested for possible use in future editions of the test. Therefore, you may not have the same test book as your neighbor. Answers to these trial questions will not be counted in your scores" (GRE National Administrations, 1977–78 *Supervisor's Manual*, p. 15). Taking the test is considered to be acceptance of or consent to that situation.

Content Characteristics

The content of each edition of the Aptitude Test is determined by concern for appropriateness to the population and comparability with past editions. Appropriateness of the material is assured by inclusion of diversified content, use of a variety of kinds of questions, and selection of nontechnical material of the sort likely to have been encountered by students planning graduate study. Comparability or stability is maintained by constant requirements for similarity in content and statistical characteristics in all editions of the test.

In the following discussion of the Aptitude Test, the part of each question that poses the problem will be referred to as the "stem," the answer choices as "options," the wrong choices as "distractors," and the right choice as the "correct response."

Verbal Ability Measure

The verbal ability measure, designed to test the ability to understand and manipulate written words in order to solve problems, consists of four question types: antonyms, analogies, sentence completions (discrete questions, so called because each question is independent, sharing no common stimulus material), and reading comprehension sets. Discrete questions are drawn from four areas of human interest: 1) the arts and humanities, 2) the social studies and concerns of practical or everyday life, 3) the world of science and nature, and 4) the domain of human relationships and feelings. Reading passages may be drawn from the humanities, social sciences, and natural sciences and may represent a narrative as well as a discursive style.

Equal amounts of time are devoted to discrete questions and sets of reading comprehension questions. Fifty-five discrete questions can be administered in 25 minutes and 25 reading comprehension questions in 25 minutes. Discrete questions are notable for their efficiency (contributing high reliability for the amount of time invested), and reading comprehension questions are distinguished by the close link they provide between the test and the actual reading activities of graduate students.

Antonyms. Antonym questions provide the least context. An isolated word or phrase is presented in the stem; the options consist of possible antonyms to the stem. Distractors may be chosen on the basis of their similarity in sound or spelling to other words, but synonyms are avoided as distractors, since they may prove more tricky than challenging to students. The purpose of antonym questions is to measure not only knowledge of words but also the ability to reason from a positive to a negative concept, to leap conceptually from one extreme to another. Word frequency lists are not generally used in selecting words to be used in antonym questions because the pretesting process provides the best indication of the familiarity of the population with the word. The difficulty of an antonym question may reflect the frequency of appearance of the words in speech and writing as well as the attractiveness of the options.

Antonyms may require only rather general knowledge of a word (see the first example below), or they may require that a student make fine distinctions (see the second example). They may appear as single words or as phrases and may be any part of speech. The directions for antonym questions and three examples appear below. An asterisk denotes the correct response.

Directions: **Each question below consists of a word printed in capital letters, followed by five words or phrases lettered A through E. Choose the lettered word or phrase that is most nearly opposite in meaning to the word in capital letters. Since some of the questions require you to histinguish fine shades of meaning, be sure to consider all the choices before deciding which one is best.**

1. CONSCRIPT: (A) mediator *(B) volunteer (C) eccentric
(D) comedian (E) villain

2. MURKY: (A) clamorous (B) complex *(C) full of light
(D) endowed with beauty (E) free from error

3. PROMULGATE: (A) distort (B) demote *(C) suppress
(D) retard (E) discourage

Analogies. Analogy questions provide somewhat more context than antonyms and require the student to recognize parallel relationships. The two words in the stem are separated by a colon suggesting that they share a relationship. Each of the options presents a pair of words, again separated by a colon to suggest a relationship between them. The relationship may be of kind, size, contiguity, or degree. Analogies may be classified as independent or overlapping.

In an independent analogy, neither of the words comprising the correct response is similar in meaning to a word in the stem (for example, COLD:CONGEALMENT::heat:incandescence, where *cold* and *heat*, through extremes of the same continuum, are no more similar in meaning than *congealment* and *incandescence*). In an overlapping analogy, one or both of the words in the correct response is suggestive of the meaning of one or both of the words in the stem (for example, METAL:DROSS::wheat:chaff, where *dross* and *chaff* both signify a waste product). There are more independent than overlapping analogies, and, even in overlapping analogies, dependence on word associations alone for the solution is avoided, often by inclusion of distractors with similar associations. The purpose of the analogy is to test the ability to recognize parallel rather than loosely related word pairs. Analogies may also be based on words with only concrete referents (see the first example below), with only abstract referents (see the second example), or with both kinds of words (see the third example). The directions for analogy questions and three examples follow. An asterisk denotes the correct response.

Directions: **In each of the following questions, a related pair of words or phrases is followed by five lettered pairs of words or phrases. Select the lettered pair which best expresses a relationship similar to that expressed in the original pair.**

1. FROND:FERN:: (A) acorn:oak (B) bulb:tulip
***(C) needle:pine (D) desert:cactus (E) foliage:blossom**

2. OBEDIENT:OBSEQUIOUS:: (A) ludicrous:ridiculous
***(B) helpful: officious (C) unusual:obvious**
(D) happy:zealous (E) serene:agitated

3. JUNTA:POLITICAL:: (A) team:successful
***(B) council:advisory (C) jury:secretive**
(D) catalogue:arbitrary (E) parent:instructive

Sentence Completions. The third discrete question type, sentence completions, provides increased context and is closely related to reading comprehension. Sentences are usually selected from reading materials that might be commonly available to students. They contain one blank or two, and students are required to select the completion that is logically and stylistically consistent with the rest of the sentence.

The four examples of sentence completions illustrate, in order, the four areas of human interest to which an antonym, analogy, or sentence completion question may be related: 1) the arts and humanities, 2) the social studies and concerns of practical or everyday life, 3) the world of science and nature, and 4) the domain of human relationships and feelings. An asterisk denotes the correct response.

Directions: **Each of the sentences below has one or more blank spaces, each blank indicating that a word has been omitted. Beneath the sentence are five lettered words or sets of words. You are to choose the one word or set of words which, when inserted in the sentence, best fits in with the meaning of the sentence as a whole.**

1. **Some time ago translators realized that they must——the idea that an ancient classic, simply because it is ancient, must be rendered in the archaic English of another era.**

   (A) extract (B) absolve (C) maintain (D) perpetuate
   ***(E) relinquish**

2. **Some people argue that the growth of industrial research has been too rapid, that in some companies research is——which is supported because of the——associated with it rather than because of the real benefits derived.**

   ***(A) a fad..glamour (B) a luxury..profit**
   (C) a necessity..satisfaction (D) an obstacle..prestige
   (E) an innovation..stability

3. **When a new comet appeared in 1577, its path straight through what were supposed to be the——spheres that formed the skies——the view that these spheres did not exist.**

   (A) solid..punctured (B) vacant..dispelled
   ***(C) impenetrable..encouraged (D) invisible..exploded**
   (E) perforated..corroborated

4. **She was saddened to hear that her colleagues continued to ——her protege, for she had hoped that success would—— him.**

   (A) patronize..enrage *(B) disparage..vindicate
   (C) underwrite..attract (D) flatter..encourage
   (E) deride..humiliate

Reading Comprehension Sets. Reading comprehension passages are of varying lengths. In each edition of the test, there are two relatively long passages, each providing the basis for answering seven or eight questions, and three relatively short passages, each providing the basis for answering three or four questions. Test forms are comparable in terms of the total number of words in passages in the reading comprehension section.

Although the mean difficulty of the questions themselves for the examinees is considered the best index of the difficulty of the reading comprehension section of the test, an attempt is made to achieve an appropriate range and variation of levels of difficulty of the reading material. In a special analysis, applying the Simple Test Approach for Readability (STAR) developed by General Motors, an average of 14.3 grade level equivalency was obtained for two recent GRE verbal forms. This grade-level equivalent suggests that the overall reading level was not difficult for college graduates. The grade-level equivalency of passages ranged from 10.1 to 21.2 in this analysis, and the correlation between mean question difficulty and the difficulty of the passage on which the questions are based was only .22. It is not surprising that the difficulty of questions has a low relationship to the difficulty of the passage associated with those questions. A question's difficulty depends on a number of factors such as the attractiveness of the distractors and the type of reading skill being tested.

The six example questions illustrate, in order, the six major types of reading comprehension questions that appear in the test. These types focus on 1) the main idea or primary purpose of the passage; 2) information explicitly stated in the passage; 3) information or ideas implied or suggested by the author; 4) possible application of the author's ideas to other situations; 5) the author's logic, reasoning, or persuasive techniques; and 6) the tone of the passage or the author's attitude as it is revealed in the language used. An asterisk denotes the correct response.

***Directions:* Each passage in this group is followed by questions based on its content. After reading a passage, choose the best answer to each question and blacken the corresponding space on the answer sheet. Answer all questions following a passage on the basis of what is *stated* or *implied* in that passage.**

The literary generation that crusaded against Puritanism and the genteel tradition, against stereotypes and sentimentality, also saw what Merle Curti terms "the beginnings of a new and realistic interest in American regions and American folk." The trend toward twentieth-century realism exemplified in this country by the works of Sherwood Anderson, Sinclair Lewis, and Theodore Dreiser was paralleled in the work of black writers of the same period. Rudolph Fisher, Arna Bontemps, and Jessie Fauset, for example, reached an almost full scale of self-revelation and a substantial degree of self-criticism. By breaking with past literary tradition, black writers in the 1920's were developing a greater sophistication of style and wider and more universal appeal.

American art faced a problem in the early twenties—a problem born of the fact that for years the white American artist had regarded the American art scene as unsophisticated, and the black artist had felt oppressed by the social situation. Frequently escaping to life abroad where new developments in art were taking place, neither contributed much toward the development of a distinctive American art. In the midtwenties, the same forces that inspired the upsurge of new, more realistic, and unapologetic talent in the other arts inspired changes in the attitude of artists. In addition, of course, the rising tide of modernism in Europe and at home encouraged young black painters to turn away from traditionalism in both subject matter and style.

Fortunately, as with the parallel movement in literature, this movement in painting did not lead the black artists into racialist art. On the contrary, it led them into the mainstream. American artists were beginning to develop Negro themes and subjects as new native American material. The older white artists had handled the Negro themes in a somewhat casual and superficial manner. For many young white artists of the twenties, blacks were the subject of careful and penetrating interpretation. The fact that young white American artists and their young black contemporaries shared this new interest in black life was significant. A common ground was established among young artists. The notion of the black world as a restricted province to which the black artist was confined was removed. At the same time, the black artist was challenged by the task of self-revelation and forced to attempt it in competition with other artists. The poise and originality of the young artists of that period and their honest depiction of American life brought them closer to the realization that race was a medium of expression, not an end in itself. For though their work was avowedly racial for the most part, they ranged with an increasing sense of freedom through the universe of a common human art. The strength and vigor of artists like Aaron Douglas, Palmer Hayden, and Hale Woodruff were a reflection of superior advantages and training. Of equal, if not greater, importance was the fact that their spiritual enlargement stemmed from the growing conception of American culture as vitally and necessarily including the materials of black life.

1. The author's primary purpose in the passage is to
   - (A) enumerate several dilemmas faced by black artists in America
   - (B) explain the differences between realism in literature and realism in painting
   - (C) contrast the works of black artists with those of their white contemporaries
   - (D) analyze the effect on black artists of the movement toward realism in art
   - (E) encourage the inclusion of black life in artistic depictions of American culture

2. The author mentions Sinclair Lewis and Jessie Fauset as examples of writers who
   - (A) awakened European interest in American culture
   - (B) broke away from past literary traditions
   - (C) portrayed the lives of blacks realistically
   - (D) were among the most prolific writers of the 1920's
   - (E) influenced artists in fields other than literature

3. It can be inferred that, in the early decades of the twentieth century, many American painters went abroad because they
   - (A) hoped to redress social injustices in America
   - (B) disliked the trend toward modernism in America
   - *(C) regarded Europe as the place where new developments in art were taking place
   - (D) wished to encourage Europeans to join the movement toward realism
   - (E) Could find no way to support themselves in America

4. The statements in the passage suggest that the author would most likely react to a movement among black artists toward racialist art with
   - (A) amused cynicism
   - (B) deliberate indifference
   - (C) enthusiastic encouragement
   - (D) cautious optimism
   - *(E) disappointed disapproval

5. The author quotes Merle Curti in order to
   - *(A) support his own analysis of a trend
   - (B) indicate the ambiguity of his topic
   - (C) provide a contrasting viewpoint
   - (D) foreshadow new directions attitudes may take
   - (E) illustrate past resentment to change

6. The tone and content of the passage suggest that its source was most likely
   - (A) a guidebook to a collection of paintings by black artists
   - *(B) an essay on the development of a characteristic American style by black and white artists
   - (C) an editorial on ghetto life as experienced by artists during the 1920's
   - (D) a book on the way art reflects public opinion as exemplified by the trend toward realism
   - (E) a biography of a famous black American artist who lives in Europe

Content Specifications for the Verbal Ability Measure

Content specifications or statements of required numbers and kinds of content for each test form assure the parallelism of all forms. The tables below and on page 14 show the breakdown of content for the current verbal ability measure.

As Tables 1, 2, and 3 illustrate, balance of diversified materials is a primary consideration to make the test appropriate for the various segments of the population. For example, since Coffman (1965) has shown that men tend to do slightly better on discrete questions classified as belonging to the world of science and nature and to the domain of social studies and concerns of practical or everyday life whereas women tend to do slightly better on the arts and humanities (aesthetic, philosophical) and human relationships questions, balance among those classes of questions should result in a test appropriate for both sexes. The rationale for balancing content throughout the test is an extension of these observations. The greater the variety of material provided, the more likely it is that the diverse population will be well served, assuming that the material is generally accessible (nontechnical).

Table 1: Specifications for Discrete Verbal Questions (55 questions)

Content	Number of Questions
Antonyms	
Arts and Humanities	5
Social Studies and Practical or Everyday Life	5
Science and Nature	5
Human Relationships and Feelings	5
General definitions	8-14
Fine distinctions	6-12
Single words	10-16
Phrases	4-10
Verb	3-9
Noun	3-9
Adjective	5-11
Other parts of speech	
Analogies	
Arts and Humanities	4-6
Social Studies and Practical or Everyday Life	4-6
Science and Nature	4-6
Human Relationships and Feelings	4-6
Concrete	4-8
Mixed	5-11
Abstract	4-8
Independent	11-15
Overlapping	5-9
Sentence Completions	
Arts and Humanities	4
Social Studies and Practical or Everyday Life	5
Science and Nature	4
Human Relationships and Feelings	4
One blank	5-9
Two blanks	11-15

21

Table 2: Specifications for Reading Comprehension Passages (5 passages)

Subject Matter	Number of Passages		
	Length		Total
	Approx. 450 words	Approx. 150 words	
Humanities	1		1
Social Studies		2	2
Natural Science	1		1
Other		1	1
Total	2	3	5

Table 3: Specifications for Reading Comprehension Questions (25 questions)

Type of Question	Number of Questions
Main Idea	3-5
Explicit Statement	5-8
Inference	6-7
Application	2-3
Logic	2-3
Tone	1-2
Total	25

Quantitative Ability Measure

The quantitative ability section is designed to test basic mathematical skills, understanding of elementary mathematical concepts, and ability to reason quantitatively and to solve problems in a quantitative setting. This section consists of three question types: discrete mathematics, data interpretation, and quantitative comparison.

Each discrete mathematics queston contains all the information needed for answering the question, except for the basic mathematical knowledge assumed to be common to the backgrounds of all students. Many of these questions, such as examples 1 and 2, require little more than manipulation and very basic knowledge; others, such as examples 3 and 4, require the student to read, understand, and solve a problem that involves either an actual or an abstract situation.

The data interpretation questions, like the reading comprehension questions in the verbal measure, usually appear in sets based on stimulus material that precedes the questions. The stimulus material for these questions consists of data presented in graphs or tables. Data interpretation questions are designed to test the ability to synthesize information, to select the appropriate data for answering a question, as in example 5, or to determine that sufficient information for answering a question is not given, as in example 6.

Directions for discrete mathematics and data interpretation questions and examples of each type follow. An asterisk denotes the correct response.

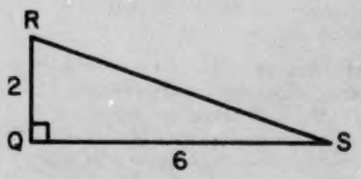
Directions: **Solve each of the following problems, using any available space on the page for scratch work. Then indicate the best answer in the appropriate space on the answer sheet.**

Note: **Figures which accompany these problems are intended to provide information useful in solving the problems. They are drawn as accurately as possible EXCEPT when it is stated in a specific problem that its figure is not drawn to scale. All figures lie in a plane unless otherwise indicated.**

All numbers used are real numbers.

1. If $\frac{3}{2}x = \frac{3}{2}$, then $1 - x =$

(A) $-\frac{3}{2}$ (B) $-\frac{5}{4}$ (C) $-\frac{1}{2}$ *(D) 0 (E) $\frac{1}{2}$



2. If A represents the area of ΔQRS above, then $2A =$

(A) 4 (B) 6 *(C) 12 (D) 24 (E) 36

3. After an initial deposit of x dollars, the amount of money in a certain fund is doubled at the end of each month for 5 months. If at the end of the 5-month period there is a total of $560 in the fund, how much money was in the fund at the beginning of the third month?

(A) $17.50 (B) $35 *(C) $70 (D) $140 (E) $224

4. Suppose that $\ominus$ stands for a binary operation which adds the reciprocals of the two numbers it operates on. For example,

$$5 \ominus 7 = \frac{1}{5} + \frac{1}{7}.$$

Which of the following statements is (are) true for all positive a,b?

I. $\frac{1}{a} \ominus \frac{1}{b} = a + b$

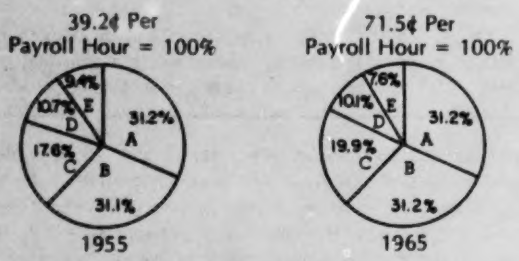
II. $a \ominus b = \frac{1}{a + b}$

III. $\frac{1}{a \ominus b} = \frac{ab}{a + b}$

(A) I only (B) II only (C) III only (D) I and II only
*(E) I and III only

Questions 5–6 refer to the following graphs.

FRINGE BENEFIT PAYMENTS BY TYPE,
1955 AND 1965, ALL INDUSTRIES



Type A. Paid vacations, holidays, sick leave, etc.
Type B. Private pension and welfare fund contributions, severance pay, etc.
Type C. Legally required payments (social security, etc.)
Type D. Paid rest periods, national guard, jury duty, etc.
Type E. Profit-sharing payments, other bonuses

5. What was the approximate increase in cents per payroll hour from 1955 to 1965 for type C fringe benefit payments?

(A) 0.9¢ (B) 1.6¢ (C) 2.3¢ *(D) 7.3¢ (E) 9.2¢

6. If fringe benefit payments averaged 25 percent of the total payroll in 1965, then fringe benefit payments averaged approximately what percent of the total payroll in 1955?

(A) 10% (B) 14% (C)25% (D) 46%
***(E) It cannot be determined from the information given.**

The third question type, quantitative comparisons, was included in the GRE Aptitude Test for the first time in the 1977–78 year, although variations of this type of question have been used for a number of years in other testing programs. Quantitative comparisons are characterized by a fixed set of four options and are the least time consuming of the three types of questions in this section; the data interpretation questions are the most time consuming. The efficiency of quantitative comparisons was one factor permitting restructuring of the Aptitude Test to include a new measure. Since performance of quantitative comparisons correlated so highly (approximately .90) with performance on other types of quantitative questions used in the test before restructuring, it was possible to reduce the testing time without reducing the number of questions. Some of the more time-consuming questions were replaced by quantitative comparisons, with the expectation that the high reliability of the test and the comparability of scores would be maintained.

Quantitative comparisons are designed to test the ability to reason quickly and accurately about the relative sizes of two quantities or to perceive that not enough information is given to make such a decision. Some questions, as in example 1 at the right, only require some manipulation to determine which of the quantities is greater—the one in Column A or the one in Column B. Other questions require the student to reason more or to think of special cases in which the relative sizes of the quantities reverse, or, as in example 2, to visualize other possible ways in which a figure could be drawn within the ground rules for figures given in the directions.

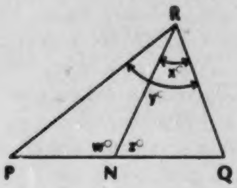
Directions: **Each of the following questions consists of two quantities, one in Column A and one in Column B. Your are to compare the two quantities and on the answer sheet blacken space**

A if the quantity in Column A is the greater;
B if the quantity in Column B is the greater;
C if the quantities are equal;
D if the relationship cannot be determined from the information given.

Common Information:	**In a question, information concerning one or both of the quantities to be compared is centered above the two columns. A symbol that appears in both columns represents the same thing in Column A as it does in Column B.**
Numbers:	**All numbers used are real numbers.**
Figures:	**Position of points, angles, regions, etc. can be assumed to be in the order. shown.** **Lines shown as straight can be assumed to be straight.** **Figures are assumed to lie in the plane unless otherwise indicated.** **Figures which accompany questions are intended to provide information useful in answering the questions. However, unless a note states that a figure is drawn to scale, you should solve these problems NOT by estimating sizes by sight or by measurement, but by using your knowledge of mathematics (see example 2 below).**

	Column A	***Column B***	***Sample Answers***
Example 1:	**2 x 6**	**2 + 6**	● Ⓑ Ⓒ Ⓓ

Examples 2-4 refer to ΔPQR.



	Column A	***Column B***	***Sample Answers***
Example 2:	**PN**	**NQ**	Ⓐ Ⓑ Ⓒ ● **(since equal measures cannot be assumed, even though PN and NQ appear equal)**
Example 3:	**x**	**y**	Ⓐ ● Ⓒ Ⓓ **(since N is between P and Q)**
Example 4:	**w + z**	**180**	Ⓐ Ⓑ ● Ⓓ **(since PQ is a straight line)**

Content Specifications for the Quantitative Ability Measure

Content specifications for the quantitative measure are given in Table 4 below:

Table 4: Quantitative Specifications for the GRE Aptitude Test

	Arithmetic	Algebra	Geometry	Total
Discrete Mathematics	5	5	5	15
Data Interpretation	8 - 10	0 - 2	0 - 1	10
Quantitative Comparisons	10	10	10	30
Total	23 - 25	15 - 17	15 - 16	55

Since this section of the test is designed primarily to measure the ability to reason quantitatively, the mathematics required does not extend beyond that assumed to be common to the mathematics background of all students. Questions classified as arithmetic can be answered by performing arithmetic operations (add, subtract, multiply, divide, find percents or averages), by reasoning, or by a combination of the two.

The algebra required does not extend beyond that usually covered in a first-year high school course and includes such topics as properties of numbers (odd and even integers, prime numbers, divisibility, and factors), operations with signed numbers, linear equations, factorable quadratic equations, factoring, simplifying algebraic expressions, exponents, and radicals. The skills required include the ability to solve simple equations, the ability to read and set up an equation for solving a complex problem, and the ability to apply basic algebraic skills to solve unfamiliar problems. Unusual notation is used only when it is explicitly defined for a particular question.

The geometry is limited primarily to measurement and intuitive geometry or spatial visualization. Topics include properties associated with parallel lines, circles, triangles, rectangles, and other polygons and measurement-related concepts of area, perimeter, volume, the Pythagorean theorem, and angle measure in degrees. Knowledge of simple coordinate geometry and special triangles such as isosceles, equilateral, and 30°–60°–90° triangles is also assumed. Knowledge of theorems and the ability to construct proofs that are usually learned in a formal geometry course are not measured.

Analytical Ability Measure

Questions in this new measure are designed to tap students abilities to recognize logical relationships (for example, between evidence and a hypothesis, between premises and a conclusion, or between stated facts and possible explanations); to judge the consistency of interrelated statements; to draw conclusions from a complex series of statements; to use a sequential procedure to eliminate incorrect choices in order to reach a conclusion; to make inferences from statements expressing relationships among abstract entities such as nonverbal or nonnumerical symbols; and to determine relationships between independent or interdependent categories or groups. Three types of questions are used in measuring these analytical skills: analysis of explanations, logical diagrams, and analytical reasoning. If continuing research should identify other question types that also effectively tap these skills, the content of the measure may gradually change if such change is demonstrated to represent an improvement. As experience with the new measure accumulates, changes may also be made in the directions, to simplify and clarify them wherever possible, or in the format. Such changes will be made under conditions designed to maintain score comparability from one test edition to another.

Analysis of Explanations. Each set of analysis of explanations questions is preceded by a narrative describing related events and a statement of a result, which may be surprising in light of the facts presented. Actually, the result may not follow directly from the situation but may be dictated by other events consistent with the situation, although not described. One part of the student's task is to imagine what missing information might plausibly explain the result. Although the measure is called analysis of explanations, its purpose may be broader than that title implies. It measures the ability to recognize inconsistencies and deducible information, to hypothesize, and to judge the relevance of certain facts to possible hypotheses or possible explanations of a stated fact. The measure also requires that a sequential procedure be followed in arriving at the correct answer. Choice A must be eliminated before choice B can be considered, and so on. Since this is a fixed-format type of question, each question in a set presents the same five answer choices. The directions and sample questions are given below.

Directions: **For each set of questions, a fact situation and a result are presented. Several numbered statements follow the result. Each statement is to be evaluated in relation to the fact situation and result.**

Consider each statement separately from the other statements. For each one, examine the following sequence of decisions, in the order A, B, C, D, E. Each decision results in selecting or eliminating a choice. *The first choice that cannot be eliminated is the correct answer.*

A Is the statement *inconsistent* with, or contradictory to, something in the fact situation, the result, or both together?
If so, choose A.
If not,

B Does the statement present a *possible adequate explanation* of the result?
If so, choose B.
If not,

C Does the statement have to be true if the fact situation and result are as stated?
If so, the statement is *deducible* from something in the fact situation, the result, or both together; choose C.
If not,

D Does the statement either support or weaken a possible explanatiion of the result?
If so, the statement is *relevant* to an explanation; choose D.

E If not, the statement is *irrelevant* to an explanation of the result; choose E.

Use common sense to decide whether explanations are adequate and whether statements are inconsistent or deducible. No formal system of logic is presupposed. Do not consider extremely unlikely or remote possibilities.

Situation: In an attempt to end the theft of books from Parkman University Library, Elnora Johnson, the chief librarian, initiated a stringent inspection program at the beginning of the fall term. At the library entrance, Johnson posted inspectors to check that each library book leaving the building had a checkout slip bearing the call number of the book, its due date, and the borrower's identification number. The library retained a carbon copy of this slip as its only record that the book had been checked out. Johnson ordered the inspectors to search for concealed library books in attaché cases, bookbags, and all other containers large enough to hold a book. Since no new personnel could be hired, all library personnel took turns serving as inspectors, though many complained of their embarrassment in conducting the searches.

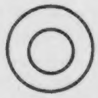
Result: During that term Margaret Zimmer stole twenty-five library books.

1. **Zimmer stole the books before the inspection system began.** (Correct response A)
2. **Zimmer dropped the books out of a second-story window into a clump of bushes and retrieved them after she left the building.** (Correct response B)
3. **During that term, if Zimmer carried a bookbag out of the library entrance door during regular hours, an inspector was supposed to check it.** (Correct response C)
4. **The doors to the library fire escapes are equipped with alarm bells set off by opening the doors.** (Correct response D)
5. **The library had at one time kept two carbon copies of each checkout slip.** (Correct response E)

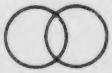
Logical Diagrams. Logical diagrams is also a fixed-format measure; that is, the same options apply to each of several sets of questions. Students are given five circle diagrams expressing different class relationships. They are then asked to look at sets of words and choose the diagram that best illustrates the relationship of the concepts they signify. The logical process might technically be described as consisting of three steps: 1) translating words into propositions that define their relationships (as in example 3 below, translating "fish, minnows, things that live in water" into the propositions that "some things live in water," and "all minnows are fish"); 2) diagramming those propositions; and 3) selecting from five diagrams the one that is appropriate to show the relationship of the propositions. The final step in the logical process—drawing the inference that "all minnows live in water"—is not required, although the diagram illustrates that inference. It should be emphasized that a student need not have studied this process formally to solve the problems; nor will the student necessarily be aware of the steps of reasoning taken to select the correct answer. The purpose is to measure skills likely to have been learned in a variety of contexts and in academic study of most kinds. The directions and sample questions follow.

Directions: **In this part, you are to choose from five diagrams the one that illustrates the relationship among the given classes better than any of the other diagrams offered.**

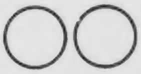
There are three possible relationships between any *two* different classes:



Indicates that **one class is completely contained in the other, but not vice versa.**



Indicates that **neither class is completely contained in the other, but the two do have members in common.**



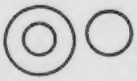
Indicates that **there are no members in common.**

Note: **The size of the circles does *not* indicate relative size of the classes.**

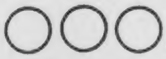
Example:

Birds, robins, trees

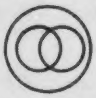
(A)



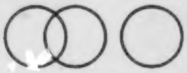
(B)



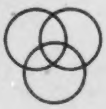
(C)



(D)



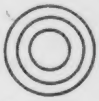
(E)



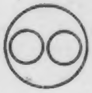
The correct answer, (A), shows that one of the classes (trees) has no members in common with the other two. (No trees are either birds or robins, and no birds or robins are trees.) (A) also shows that one of the two remaining classes (robins) is completely included in the other class (birds).

The five possible choices for all problems in this part are given below.

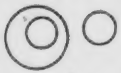
(A)



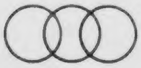
(B)



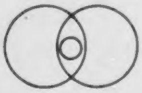
(C)



(D)



(E)



1. **Nuts, pecans, forks** (Correct response C)
2. **Adult women, infants, black-haired people** (Correct response D)
3. **Fish, minnows, things that live in water** (Correct response A)

Analytical Reasoning. Analytical reasoning consists of complex sets of statements from which the student must draw inferences. The statements may include abstractions such as symbols without specific referents. The directions and sample questions appear below. An asterisk denotes the correct response.

Directions: **Each question or group of questions is based on a passage or set of statements. In answering some of the questions it may be useful to draw a rough diagram. Choose the best answer for each question and blacken the corresponding space on your answer sheet.**

Questions 1–2

(1) It is assumed that a half tone is the smallest possible interval between notes.
(2) Note T is a half tone higher than note V.
(3) Note V is a whole tone higher than note W.
(4) Note W is a half tone lower than note X.
(5) Note X is a whole tone lower than note T.
(6) Note Y is a whole tone lower than note W.

1. Which of the following represents the relative order of the notes from the lowest to the highest?

(A) X Y W V T   *(B) Y W X V T   (C) W V T Y X
(D) Y W V T X   (E) Y X W V T

2. Which of the following statements about an additional note, Z, could NOT be true?

(A) Z is higher than T.   (B) Z is lower than Y.
(C) Z is lower than W.   (D) Z is between W and Y.
***(E) Z is between W and X.**

Questions 3–4

(1) You cannot enter unless you have a red ticket.
(2) If you present a blue form signed by the director, you will receive a red ticket.
(3) The director will sign and give you a blue form if and only if you surrender your yellow pass to him.
(4) If you have a green slip, you can exchange it for a yellow pass, but you can do so only if you also have a blue form signed by the director.
(5) In order to get a red ticket, a person who does not have a driver's license must have a blue form signed by the director.
(6) You can get a yellow pass on request, but you can do so only if you have never had a green slip.

3. The above procedures fail to specify

***(A) whether anything besides a red ticket is required for entrance**
(B) whether you can exchange a green slip for a yellow pass
(C) the condition under which the director will sign the blue form
(D) how to get a red ticket if you have a yellow pass
(E) whether it is possible to obtain a red ticket if you do not have a driver's license

4. Which of the following people can, under the rules given, eventually obtain a ticket?

I. A person who has no driver's license and who has only a green slip
II. A person who has no driver's license and who has only a yellow pass
III. A person who has both a driver's license and a blue form signed by the director

(A) I only   (B) II only   (C) I and II only
***(D) II and III only   (E) I, II, and III**

Content Specifications for the Analytical Ability Measure

The content specifications for the analytical ability measure are based on achieving an approximate balance between questions with greater face validity for students with a humanities-social-studies orientation and those with greater face validity for students with a science orientation (though those categories are clearly not exclusive or independent). The specifications now call for 40 analysis of explanations questions (appearing more closely related to kinds of analysis used in the humanities and social studies) and 30 questions (15 logical diagrams and 15 analytical reasoning) appearing more closely related to the kind of analysis required in the sciences. Since the analytical measure was introduced for the first time in October of 1977, a detailed breakdown of specifications is not yet in final form. Diversity of subject matter and questions is the general rule.

Statistical Characteristics

Item and test analyses, which are regularly performed for every new test form introduced, provide information on the statistical characteristics of the test and its components. The most important statistical information indicates the difficulty, reliability, interrelationship of test components, and speededness. Data on these characteristics for five recent Aptitude Test forms administered prior to October 1977 are shown in Table 5, and data for the first two restructured test forms administered in October 1977 are shown in Table 5A. The analyses providing these data were based on samples representative of the administrations in which the respective tests were introduced rather than the total GRE population in a given year or years. Of the five prior forms shown in Table 5, three were introduced in April and two in October. The October examinees are consistently more able, on the average, than the April examinees.

The reliability estimate for the verbal sections of a typical prior form of the Aptitude Test is .93 and for the quantitative section .91, with corresponding standard errors of measurement of 33 and 40 converted scaled score values, respectively.* Taken separately, the two verbal components—discrete verbal questions and reading comprehension—have reliabilities in the middle to upper .80s. Thus, the reliability of each of the verbal sections is higher than the intercorrelation between the two, which is in the middle .70s, suggesting that the two verbal components contribute somewhat independent indicators of verbal ability.

The correlation between the verbal ability score and the quantitative ability score is about .56 in a typical prior form. The correlation between the discrete verbal and quantitative sections is .46 and the

---

*Reliability is estimated by the Kuder-Richardson formula (20), adapted for use with formula scoring.

Table 5: Statistical Characteristics of Five Recent Prior Forms of the GRE Aptitude Test

Time Introduced	Statistics Based on Total Groups					Statistics Based on Evenly Spaced Samples of Total Groups											
	Number of Examinees	Range of Possible Scaled Scores		Mean and Standard Deviation of Scaled Scores		Number of Examinees	Average Percentage of Examinees Answering Questions Correctly		Reliability		Correlation between V & Q	Percentage of Examinees Completing: ¾ of Test Section			Entire Test Section		
												V*		Q	V*		Q
		V	Q	V	Q		V	Q	V	Q		Sec. I	Sec. II		Sec. I	Sec. II	
April 1973	26,096	200-830	200-830	485 SD 115	490 SD 133	1,090	53%	57%	.93	.91	.56	94%	98%	97%	55%	78%	45%
October 1973	33,427	220-840	200-850	519 SD 120	519 SD 137	1,115	56%	59%	.93	.90	.53	94%	97%	97%	55%	71%	50%
April 1974	27,287	200-850	200-820	475 SD 127	489 SD 134	1,130	51%	57%	.93	.91	.58	92%	97%	96%	52%	76%	44%
April 1975	25,978	200-860	200-820	476 SD 126	486 SD 134	1,600	55%	57%	.93	.91	.56	94%	98%	96%	58%	78%	49%
October 1976	29,229	200-860	200-840	512 SD 131	525 SD 137	1,960	59%	63%	.94	.91	.60	96%	98%	96%	66%	84%	36%

*The verbal questions of the prior Aptitude Test were in two separately timed sections. Section I contained discrete questions and was given in 25 minutes. Section II contained reading comprehension questions and was given in 50 minutes.

Table 5A: Statistical Characteristics of the First Two Restructured Forms of the GRE Aptitude Test

(Based on Two Separate Evenly Spaced Samples of 1,950 and 1,945
Examinees at the October 1977 Administration)

Form	Range of Possible Scaled Scores			Mean and Standard Deviation of Scaled Scores			Average Percentage of Examinees Answering Questions Correctly			Reliability			Correlation between			Percentage of Examinees Completing: ¾ of Test Section				Entire Test Section			
																V	Q	A*		V	Q	A*	
	V	Q	A	V	Q	A	V	Q	A	V	Q	A	V & Q	V & A	Q & A			Sec. III	Sec. IV			Sec. III	Sec. IV
1	220-850	200-880	210-810	501 SD 126	526 SD 133	510 SD 131	54%	60%	57%	.93	.88	.92	.57	.77	.74	88%	98%	95%	97%	31%	56%	73%	60%
2	210-850	200-870	210-800	502 SD 125	525 SD 132	515 SD 129	55%	62%	62%	.93	.88	.92	.54	.76	.71	87%	99%	96%	98%	24%	56%	79%	63%

*The analytical questions of the restructured Aptitude Test are in two separately timed sections. Section III contains analysis-of-explanations questions and is given in 25 minutes. Section IV contains logical diagrams and analytical reasoning questions and is given in 25 minutes.

correlation between the reading comprehension and quantitative sections is .59.

For the first two forms of the present restructured Aptitude Test administered in October 1977, the reliability of the verbal ability measure, which has discrete verbal and reading comprehension questions combined in one section, remains at .93. The reliability of the shortened quantitative ability section is .88, and that for the new analytical ability measure .92. Standard errors of measurement for the scores on the restructured test are 33 for verbal ability, 38 for quantitative ability, and 36 for analytical ability.

The correlation between the verbal ability and quantitative ability scores on the restructured test is approximately .54; between the verbal ability and analytical ability scores .73; and between the quantitative ability and analytical ability scores .70.

One set of standards sometimes taken to indicate that a test is a power test and lacks any significant speed factor is that virtually all examinees reach three-fourths of the questions and 80 percent reach the last question. The percentage completing three-fourths of the test is the more reliable indicator because the percentage completing the test depends entirely on the number answering the very last question. Often there is quite a large difference between the percentage reaching the next-to-last question and the percentage reaching the last question.

In terms of this set of standards, the test sections of five recent forms of the prior Aptitude Test were slightly speeded for those taking them. The percentages completing three-fourths of the test sections ranged from 92 to 98 percent with a median value of 96 percent. The verbal sections of the first two restructured forms are somewhat speeded, but the other sections are only slightly speeded.

Another approach to investigating speededness is factor analysis. Factor analyses were performed on two forms of the GRE Aptitude Test given in October 1975 (Powers, Swinton, & Carlson, 1977). The results showed that speededness associated with the discrete verbal questions accounted for 6.2 and 7.7 percent of the common variance of the first and second forms, respectively, whereas the factor reflecting speededness in the reading comprehension passages explained only 2.5 and 4.4 percent. A factor of quantitative speed, accounting for 2.5 percent of the common variance, emerged as a separate factor in the second form only.

Statistical Specifications

The statistical specifications for each form of the test are fairly constant and change only gradually and for compelling reasons. The purpose of such stability in specifications is to assure parallel forms of the test and thus comparability of scores regardless of form. Statistical adjustments for remaining unavoidable differences in test forms are thus smaller and less susceptible to error than if forms were widely divergent.

Essential to the effectiveness of detailed, fixed specifications is the pretesting process. Since all questions used in the Aptitude Test are tried out experimentally on the regular GRE population, without being counted toward students' scores, the statistical characteristics of individual questions are known and can be used in selecting questions that will meet the statistical specifications and result in an appropriate test for the population.

The primary statistical specifications are: 1) difficulty of the test (expressed as a mean delta for questions), 2) range of question difficulties, and 3) mean question-test correlation (r-biserial). The test assembler knows the difficulty and question-test correlation of each question in the usable pool. Thus, the test can be constructed to provide a full range of scores, rather than measure at one end or at the middle of the scale only, and appropriate total reliability. The statistical specifications for the Aptitude Test appear in Table 6.

Table 6: Statistical Specifications for the GRE Aptitude Test

	Mean Delta	Standard Deviation of Deltas	Mean r-Biserial
Verbal	12.3 - 12.7	2.5 - 2.8	.43 - .47
Quantitative	12.3 - 12.7	2.8 - 3.0	.50 - .55*
Analytical (tentative)	12.3 - 12.7	2.5 - 2.8	.43 - .47

*Since there are fewer quantitative questions than verbal and analytical questions, the mean *r*-biserial must be higher to obtain appropriate reliability.

Although these specifications do not provide separate requirements for each type of question within a measure, wide variations in the statistical characteristics of the components of the test are not tolerated. For example, the reading comprehension questions cannot have a mean delta of 8 while the mean delta of the discrete verbal questions is 16. The mean delta range considered acceptable for the various types of questions is approximately 11 to 13. The specifications are not always met precisely for a given administration in which the test is analyzed because the populations taking the test at different times of the year vary. This variation affects results of the analysis of a given final form and of pretest material.

Relationship of Statistical Analysis and Research to Test Specifications

Research and statistical analysis play a major role both in setting specifications for the Aptitude Test and in meeting specifications for a final form with pretested and statistically analyzed questions. Research that is closely related to the Aptitude Test is not usually formally reported or published because its primary audience is test developers who can act on the results. In some cases, materials intended for use in a future form are the subject of the research and would be compromised by publication. However, such research is an important part of the test-making process.

Standard Activities

After every form of the test is introduced, it is analyzed statistically. In addition, item (question) analyses are performed giving information on each question in the test. Item analyses for the pretest are particularly important because they enable the test specialist to identify and eliminate questions that are not performing consistently with similar questions in the final form or with the rest of those in the pretest section. Pretesting and the item analysis process provide the necessary data for meeting the specifications for a final form of the test.

Statistical information is also necessary for revising specifications where necessary. For example, in test analyses performed prior to 1972, it was noted that Section I of the verbal measure appeared to be speeded; that is, not enough people were reaching the last question, too few people were reaching three-fourths of the questions, and the variance of questions not reached was too high relative to the variance of scores. For this reason the test specifications were changed in 1972 from requiring 60 discrete verbal questions in that section to 55. Since the reliability of the verbal measure was well above .90, reliability was not threatened by this change. A gradual decrease in the number of word problems in the quantitative section of the test (that is, a reduction in the number of questions requiring a great deal of reading to understand and then solve the problem) is due to careful investigation of the correlation of the verbal and quantitative measures as well as the philosophical stance that the verbal and quantitative measures should be as independent as possible.

Special Activities

Periodically, other kinds of research or statistical analyses are carried out to evaluate or explore the possibility of changing test specifications. Analyses carried out in other programs at ETS (such as the Law School Admission Test Program, the Graduate Management Admission Test Program, and the College Entrance Examination Board's Admission Testing Program) may also contribute to the GRE test development process. Examples of analyses for other programs that have had some influence on the thinking of test developers working on the GRE are: 1) criterion validity research done by the Law School Admission Test Program on a type of question considered by the GRE for the newly restructured Aptitude Test; 2) factor analyses of the Scholastic Aptitude Test (which until very recently had content similar to the GRE Aptitude Test, though not as difficult); and 3) coaching studies carried out by the College

Board and other test sponsors to determine whether aptitude test questions are, indeed, measuring skills developed over a long period rather than skills that can be learned in a brief cramming session.

Coaching studies in a variety of settings and for a number of groups have shown that special cramming sessions or study of test-taking strategies for aptitude tests cannot substantially improve performance on quantitative and verbal questions like those that have traditionally been part of the GRE Aptitude Test. However, in a recent SAT study (Pike and Evans, 1972) in which coaching was redefined to suggest a substantial component of instruction in mathematics as well, performance on the quantitative questions—quantitative comparisons in particular—was improved. A similar study was initiated to test the findings in a GRE context, but the results were not interpretable because of the scanty data resulting from a high dropout among subjects in the experiment. Nevertheless, the SAT findings suggest that significant instruction in quantitative skills, as opposed to instruction primarily in test-taking strategies, can be effective and may be reflected in test scores. It cannot be concluded that quantitative comparisons are in the usual sense coachable since coaching and mathematics instruction were combined in the Pike and Evans study showing score improvement on that type of question. The GRE *Information Bulletin* contains sample quantitative comparisons, and a full complement of quantitative comparisons is included in the GRE Sample Aptitude Test, both of which are accessible to all students. It is assumed that any possible coaching effect—if such an effect should exist separately from instruction—will be standard for all students if all become familiar with the question type before taking the test. Future research is expected to explore further the question of coachability and instruction in the widely applicable skills measured by the GRE Aptitude Test.

The Sample Aptitude Test was first published in 1975–76 to give all students equal access to information on the kinds of questions in the test and equal opportunity to become familiar with them and ways to solve them before taking the test. Another reason for providing the Sample Aptitude Test was to make it possible for students to obtain more information on the test without turning to marketed materials that, though designed to prepare students for the GRE, may not actually parallel the content of the GRE.

Research on subpopulations has been done in the GRE Program and in other testing programs to determine the interaction of test content and students' performance. For example, in one study the performance of males, females, blacks, and whites has been analyzed to determine whether different kinds of questions or topical material have differential effects on the question difficulties for those various groups. (See "Population Validity" in Chapter 6.)

Still other special subpopulation studies concern appropriateness of the timing and directions of the test. Currently, a study is underway to determine whether allowing additional time for verbal and quantitative questions will have differential effects on subgroups of the population identified by age, sex, or ethnic characteristics. Research is also underway to examine closely the guessing procedures students use, students' attitudes toward guessing as engendered by the test instructions, and the possible differential effects of various guessing instructions on different subgroups of the population. If results of this research show that formula scoring or guessing instructions may be working to the disadvantage of some students, the method of scoring or the wording of the directions will be reconsidered.

Research Related to Restructuring the Aptitude Test

Perhaps the relationship between statistical analysis and research and the Aptitude Test development process is best illustrated by the research effort preceding the decision to restructure the Aptitude Test to include shortened versions of the verbal and quantitative measures and a new analytical measure. Because of the significance of the proposed change, the Graduate Record Examinations Board, particularly the Research Committee, monitored all stages of the research, determining what questions seemed to merit further investigation and making final decisions concerning results. The extensive research effort was focused on possible ways of broadening the definition of talent measured by the GRE Aptitude Test.

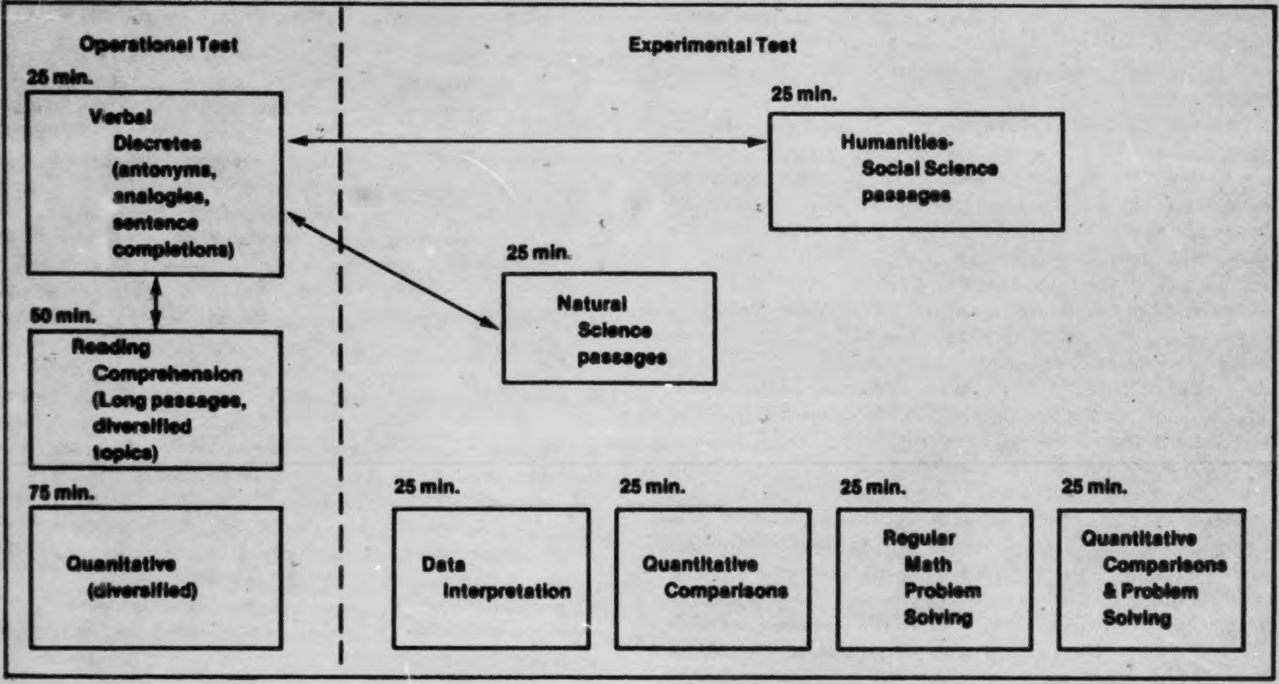
The verbal and quantitative ability measures, both of which were respected for the usefulness they had demonstrated over the years, were examined to see whether they could be made more useful and whether they could be shortened to make room for possible new measures, should they become available. Concurrently, research was undertaken to determine whether a supplemental measure could be designed that would allow students to demonstrate a broader range of skills and permit educational institutions to better judge the academic qualifications of their applicants.

Several methods of study were used: constituency surveys (questionnaires addressed to students, faculty members, and administrators), experimental pretesting followed by item and test analyses, factor analyses, validity research using self-reported undergraduate grades as the criterion, and some analyses for special subgroups of the population.

Research Related to Changes in the Verbal and Quantitative Measures. Research related to the verbal and quantitative ability measures was intended to show whether these measures could be shortened to make room for a new measure and whether the diversified content of the reading comprehension measure could be replaced by specialized material to be selected by students on the basis of their undergraduate background. It was hoped that specialized reading material would increase the validity of the verbal measure and provide a useful subscore without affecting the comparability of total verbal scores. The investigation focused on the following questions.

1. Can the GRE verbal and quantitative ability measures be shortened?
    - (a) What effect will a reduction in the number of reading comprehension questions have?
    - (b) What effect will the introduction of relatively short reading comprehension passages have?
    - (c) What effect will the introduction of quantitative comparison questions have on the quantitative measure?
2. Can reading comprehension subscores be based on different reading selections for students with different undergraduate majors? Are total verbal scores, based on reading material corresponding to undergraduate background, comparable to past reported total verbal scores based on diversified reading material?

A related subsequent question was: If it is not feasible to provide subscores based on material selected by students according to major field, can a reading comprehension subscore based on com-


Operational Test
25 min.
Verbal Discretes (antonyms, analogies, sentence completions)
50 min.
Reading Comprehension (Long passages, diversified topics)
75 min.
Quantitative (diversified)
Experimental Test
25 min.
Humanities-Social Science passages
25 min.
Natural Science passages
25 min.
Data Interpretation
25 min.
Quantitative Comparisons
25 min.
Regular Math Problem Solving
25 min.
Quantitative Comparisons & Problem Solving

mon material for all students be provided? Because of the high intercorrelation of such a subscore with the total verbal score, the answer to this question was "no."

Several methods of investigation were used. A survey of departmental representatives and administrators was made in a special questionnaire in the GRE Board *Newsletter.* Short experimental variations of the reading comprehension and quantitative measures were developed and included as trial material in a regular administration of the GRE Aptitude Test. It was then possible to compare 1) the combination of operational discrete verbal questions and the 50-minute operational section of reading comprehension questions with an experimental combination of the operational discrete verbal section and the 25-minute experimental reading comprehension section, and 2) the 75-minute quantitative measure containing three types of questions with each of four 25-minute experimental sections. The comparisons are diagrammed above. Factor analyses were performed to test the potential usefulness of the experimental material.

Research results suggested that the verbal and quantitative measures could be shortened without altering the original functions of these measures or the comparability of scores on the original and new versions. The verbal measure could contain 25 instead of 40 reading comprehension questions without falling below .90 in reliability. Because of the lower proportion of reading comprehension questions, which have a higher correlation with quantitative scores than discrete verbal questions, the separateness of the verbal and quantitative ability measures would, in fact, be enhanced. However, optional reading comprehension sections based only on specialized topical material would result in lack of equivalence between total verbal scores on the original test and the new test. The inclusion in the quantitative section of a number of quantitative comparisons would not noticeably alter the factor structure of that measure or its comparability to the original test.

Research Related to the New Analytical Measure. At the GRE Board's direction, seven types of questions intended to measure various aspects of reasoning ability were developed. Various sources of questions purported to measure reasoning or analytical skills were examined, such as some of the measures in the French Factor Kit, components of the Law School Admission Test and Graduate Management Admission Test, the Watson-Glaser Test of Critical Thinking, and the Cornell Test of Critical Thinking. However, the emphasis was on creating new question types, each intended to tap a different aspect of reasoning or analytical skills. When pretests of each of the seven question types administered with the regular GRE Aptitude Test were analyzed, a number of questions were posed: 1) Will the new test questions yield material that is appropriately difficult, reliable, and unspeeded? 2) Will they measure skills that are relatively independent of verbal and quantitative abilities? 3) Will they be valid in relation to the criterion of self-reported undergraduate grades? 4) What combination of the new test questions, if any, would be appropriate to create a new measure that would add to the value of the Aptitude Test?

To provide answers to these questions, each type of question was pretested in one of three regularly scheduled GRE national administrations. Each pretest was taken by a substantial number of GRE examinees. For all but one question type, at least three samples were drawn: a representative (spaced) sample of all students, a sample of biological and physical science undergraduate majors, and a sample of humanities and social science undergraduate majors. In addition, separate analyses for one pretest were based on samples of black males, black females, white males, and white females.

The efficiency, criterion validity, difficulty, reliability, speededness, correlation with verbal and quantitative ability measures, and appropriateness for students with different academic backgrounds were investigated for each type of question. To assess the face

Table 7: A Comparison of Various Experimental Question Types

Administration Date	Question Type	Reliability[1]	Difficulty[2]	Efficiency	Face Validity		Criterion Validity	Correlations[6] with the	
				Time Requirement	For Importance	For Measuring Abstract Reasoning		Verbal Score	Quantitative Score
December 1975	Letter Sets	.92	7.9	.7 min per question	45%	73%	.22	.48	.65
1971 and December 1975	Logical Reasoning	.68[3]	12-13	1.2 min per question	66%	67%	.25	.90	.68
1973 and December 1975	*Analytical Reasoning	.69[3]	12-13	1.2 min per question	50%	79%	.25	.78	.78
December 1975	Evaluation of Evidence	.76	12.8	.6 min per question	57%	76%	.22	.73	.59
December 1975	*Analysis of Explanations	.78	13.3	.6 min per question	57%	71%	.27	.73	.66
January 1976	*Logical Diagrams	.92	11.6	.5 min per question	52%	83%	.16-.18	.67	.77
June 1976	Deductive Reasoning	.67[4]	11.8	1.8 min per question	[5]		.13	.52	.79

[1] Estimated by the Kuder-Richardson formula (20) adapted for use with formula scoring.

[2] Difficulty is given in terms of the delta scale, with a mean of 13 and a standard deviation of 4. For five-choice questions such as these, middle difficulty is a delta 12 (60% answered correctly, 50% theoretically "knew" the answer, and 10% guessed correctly). (See also page 28 and Chapter 3.)

[3] Reliabilities for logical reasoning and analytical reasoning have been adjusted so that they are comparable to the other reliabilities, based on 25-minute sections.

[4] This reliability figure is inflated by speededness.

[5] No student questionnaire data are available.

[6] Corrected for attenuation.

validity of each question type and the way in which different groups perceived its utility, surveys were administered to samples of students who had taken each pretest, and two student committees offered opinions about samples of the experimental questions. Presentations were made at a number of national and regional meetings of professional associations, and the questions were briefly discussed by some GRE Advanced Test committees of examiners. As a result, a number of decisions could be made about the appropriateness of each of the seven question types as a possible part of a new measure.

Statistical characteristics of each type of question are indicated in Table 7. The asterisked question types are included in the new analytical measure.

References

Coffman, W. E. Principles of developing tests for the culturally different. *Proceedings of the 1964 Invitational Conference on Testing Problems.* Princeton, N.J.: Educational Testing Service, 1965.

Educational Testing Service. *A confidential testing service for use by colleges and universities in programs of appraisal, selection, and guidance* (The Graduate Record Examination General Bulletin Number 2). New York: Educational Testing Service, October 1948.

Educational Testing Service. *GRE National Administrations, 1977–78 Supervisor's Manual.* Princeton, N.J.: Educational Testing Service, 1977.

Graduate Record Examinations Board. *Newsletter.* Princeton, N.J.: Educational Testing Service, September–October 1975.

Pike, L. W., & Evans, F. R. *Effects of special instruction for three kinds of mathematics aptitude items* (CEEB Research Report 1). New York: College Entrance Examinations Board, 1972. (Monograph)

Powers, D. E., Swinton, S. S., & Carlson, A. B. *A factor analytic study of the GRE Aptitude Test* (GRE Board Professional Report 75-11P). Princeton, N.J.: Educational Testing Service, August 1977.

Chapter 4

DEVELOPMENT OF THE ADVANCED TESTS

Applicants for graduate study can demonstrate some of their qualifications by taking an Advanced Test in the discipline they studied as undergraduates. Beginning in October 1976, there were Advanced Tests in each of 20 disciplines: biology, chemistry, computer science, economics, education, engineering, French, geography, geology, German, history, literature in English, mathematics, music, philosophy, physics, political science, psychology, sociology, and Spanish.

Decisions concerning the appropriateness of providing a test are based on criteria developed by the GRE Board in 1967. These criteria, given below, generally reflect the primary purpose of an Advanced Test: to provide a high quality measure of undergraduate achievement to aid in assessing the preparedness of students for graduate study in a major discipline.

1. A significant number of institutions should offer a graduate program in the field under consideration.
2. There should be a significant number of qualified graduate faculty members whose teaching and research activities are in that field primarily.
3. A significant number of matriculated graduate students should be studying in that field primarily.
4. There should exist one or more appropriate learned societies or professional associations in that field which publish one or more scholarly journals in which original research articles are published. One or more appropriate learned societies or professional associations should express an interest in the establishment of any new test offering and a willingness to cooperate in its development.
5. The field should be sufficiently homogeneous so that a test with satisfactory psychometric characteristics can be developed.
6. There should be good reason to think that continued validation of the test by appropriate methods will yield satisfactory results.
7. The field should be amenable to techniques of testing that can be reliably scored.
8. There should be a sufficient number of potential candidates in any proposed new field so that adequate statistical data (e.g., scaling, norming) can be obtained on the first administration of a test in that field.
9. A test in the field should be amenable to standardized test administration procedures.
10. The field should be amenable to testing techniques which are comparable in spirit and quality to those used in the other Advanced Tests.
11. Introduction or continuation of a test in any one field should not impose an undue financial strain on the GRE National Program as a whole.
12. There should be a demonstrable need for a test in the field which is not met adequately by other available instruments or testing programs.

The Advanced Tests are constructed on the assumption, borne out by evidence from validity studies, that graduate school performance is related to achievement at the undergraduate level. Thus, the tests focus on measurement of learning in undergraduate curricula. Because the tests are intended primarily for graduate applicants, on the average a more able population than all undergraduate majors in a field, the tests in some cases may be relatively difficult for the average senior who does not plan to continue study. However, the tests are designed to cover the material that would be encountered by the average senior majoring in a field. Students who move from one undergraduate field to another graduate field will tend to find the test in their undergraduate major field to be more appropriate than the test offered in the field they plan to enter.

One of the main advantages of the tests is the standard measure of competence they provide. The scores reflect the relative standing of all students on the same measure. A second advantage is that subscores, reported for 9 of the 20 tests, show strengths and weaknesses in particular subfields of the disciplines.

The Advanced Tests also have limitations. Students who specialize very early and who do not have a broad background in the field may find the coverage of a test inappropriate. The tests must focus on topics to which the majority of students in a field have been exposed. In some fields, which incorporate a wide variety of relatively independent subfields—such as education and engineering, for example—it is a particular challenge to find the "core" of knowledge that is common to all.

Uses

GRE Advanced Test scores generally are used to assist in making decisions on admission to graduate programs and in awarding fellowships. The total scores serve this purpose and some subscores are sufficiently reliable (near .90) to be used in making admission decisions as well. The subscores are especially useful in counseling admitted students and helping them decide what courses they should take. Other uses of the Advanced Tests are as indicators of the effectiveness of an undergraduate or master's program and as comprehensive examinations at the undergraduate level. (The "Guidelines for the Use of the GRE" in the *Guide to the Use of the Graduate Record Examinations* includes a list of appropriate uses of the Advanced Tests.) To properly use the tests, it is important that the test content be reviewed, pertinent information in addition to test scores be considered, and the relationships between measures of the qualifications of students for graduate study and measures of their later success in graduate study be determined and recorded on a continuing basis (see the discussion of validity in Chapter 6).

Some graduate departments require Advanced Test scores, others recommend them, and still others recommend them under certain circumstances. The number and percentage of graduate

Table 8: Policies of Graduate School Departments Listed in the *Graduate Programs and Admissions Manual* on Use of the GRE Advanced Tests for the 20 Fields of Study Whose Names Match or Closely Match Those of the Tests

		Require Advanced Test		Recommend Advanced Test		Recommend Advanced Test in Certain Cases		None of the Foregoing		Total
Test	Field	No.	%	No.	%	No.	%	No.	%	No.
Biology	Biology	160	50	39	12	31	10	89	28	319
Chemistry	Chemistry	76	24	48	15	84	27	103	33	311
Computer Science	Computer Science	New—no data								
Economics	Economics	75	36	15	7	40	19	77	37	207
Education	Education, General	15	12	11	8	18	14	86	66	130
Engineering	Engineering*	77	12	136	22	163	26	255	40	631
French	French	49	37	18	14	19	14	46	35	132
Geography	Geography	20	16	14	11	26	20	68	53	128
Geology	Geosciences	53	32	47	28	16	10	51	30	167
German	German	24	22	11	10	29	26	47	42	111
History	History	101	32	42	13	49	15	127	40	319
Literature in English	English	133	39	40	12	40	12	132	38	345
Mathematics	Mathematics	69	22	43	13	78	24	129	40	319
Music	Music	54	24	16	7	24	11	129	58	223
Philosophy	Philosophy	28	19	13	9	47	31	62	41	150
Physics	Physics	81	32	25	10	79	32	65	26	250
Political Science	Political Science	67	30	26	12	48	21	84	37	225
Psychology	Psychology	143	48	26	9	39	13	88	30	296
Sociology	Sociology	63	30	21	10	47	22	82	38	213
Spanish	Spanish	13	19	11	16	11	16	35	50	70

*Includes general, chemical, civil, electrical, industrial, and mechanical engineering.

school departments requiring or recommending that students provide GRE scores in the 20 fields of graduate study for which Advanced Tests are offered are shown in Table 8. These data are taken from the *Graduate Programs and Admissions Manual*, 1976–77 edition.*

Format

All the questions in each Advanced Test are of the multiple-choice type. Eighteen tests have five-choice questions exclusively; the Advanced German Test has four-choice questions exclusively; and the Advanced Spanish Test has both four- and five-choice questions.

Various kinds of multiple-choice questions appear in the tests. Most are independent items in which a question or incomplete statement is followed by five (or four, as mentioned earlier) suggested answers or completions. The examinee is to choose the one option that best answers the question or completes the statement. Many tests contain sets of questions in which all questions in the set relate to one topic; ordinarily, a body of information on the topic is presented at the beginning of the set. Sets can lead to probing more deeply into a topic than is generally possible with independent questions. Extensive information cannot be presented with each individual question if a test is to be unspeeded. As many as 10 questions are contained in some sets, although 3 to 5 questions are more usual.

In some sets of questions, five preset answer choices in the form of words, sentences, graphs, equations, diagrams, charts, or symbols are presented and serve as the options for several questions that follow. These questions can usually be answered more rapidly than independent questions because the same five options are used repeatedly. The five options must be chosen in such a way that they possess a reasonable degree of relatedness, and all five have some plausibility as answers to each question in the set. Yet they cannot be so closely related as to be nearly synonymous, thus leading to two or more correct answers to a given question.

Some questions require the examinee to select the incorrect, least likely, or exceptional response. The nature of such a task, which is the reverse of the usual pattern, is communicated with emphasis by capitalizing words such as INCORRECT, LEAST, or EXCEPT in the stem of the questions. There is some evidence that such questions are more difficult than those in the usual pattern. It does seem important to be able to reason toward the identification of exceptions as well as correct answers, and some questions lend themselves more comfortably to this format than to the standard pattern.

---

*This edition of the *Manual* contains information about 522 institutions in the United States and about the graduate programs they offer in 79 major fields. The 522 institutions represented in the *Manual* are attended by more than 85 percent of all graduate students in the United States. The approximately 700 accredited institutions in the United States that offer master's or higher degrees in the fields of education and the liberal arts and sciences were invited to supply information for the *Manual*.

A number of questions in the tests are arranged to allow for multiple responses within the framework of recording only one mark on the answer sheet for each question. A question of this type may take the following form:

Factors responsible for the observed increase include which of the following?

I. Higher pressure
II. Higher temperature
III. Lower humidity
IV. Lower wind speed

(A) I only (B) I and II only (C) III and IV only
(D) I, II, and III only (E) I, II, III, and IV

Each Advanced Test is given in 170 minutes. All the tests are designed to be power rather than speed tests; consequently, the guideline has been adopted that virtually all examinees should complete three-fourths of the questions and about 80 percent should finish the test. Other indicators of possible speededness that are found and studied for each test edition include the means and variances of the number of questions omitted and the number of questions not reached and plots of the scores as a function of the number of right plus wrong responses. The number of questions in the tests varies from 66 (2.58 minutes per question) in the Advanced Mathematics Test to 230 (0.74 minute per question) in the Advanced Literature in English Test. In the more quantitative fields where the test questions sometimes require preliminary figuring before an answer can be selected, more time is needed per question. A complete listing of the number of questions and the average testing time per question for each of the 20 tests is given in Table 9.

Table 9: Number of Questions in GRE Advanced Tests and Average Testing Time per Question (Each test is given in 170 minutes.)

Test	Number of Questions	Average Testing Time per Question in Minutes
Biology	210	0.81
Chemistry	150	1.13
Computer Science	80	2.12
Economics	160	1.06
Education	200	0.85
Engineering	150	1.13
French	190	0.90
Geography	200	0.85
Geology	200	0.85
German	210	0.81
History	190	0.90
Literature in English	230	0.74
Mathematics	66	2.58
Music	200	0.85
Philosophy	160	1.06
Physics	90	1.89
Political Science	170	1.00
Psychology	200	0.85
Sociology	200	0.85
Spanish	210	0.81

Committees of Examiners

For each GRE Advanced Test there is a committee of examiners composed usually of five scholars in the discipline of the test. Some committees have more than five members on occasion, and some have fewer than five on much rarer occasions, but five is the typical number. One of these scholars serves as the chairperson of the committee.

The committee members or examiners are generally appointed for two-year terms. Members may be reappointed for any number of terms. In general, the membership of each committee changes gradually every two years. It is typical for one or two examiners to leave a committee and be replaced by one or two scholars new to the committee every other year. Hence, a typical length of service for a given committee member is four to eight years. The advantage of this gradual change in the membership of the committee is that it provides continuity and experience on the one hand and fresh insights and approaches on the other.

The scholars who serve on a committee of examiners are almost always college and university professors in the discipline of the test. The few exceptions to this pattern invariably arise because an examiner who was a professor at the time of the initial appointment leaves the professional ranks for some other assignment during the term of the appointment. The examiners tend to be members of graduate school faculties at universities with large graduate schools and high quality programs. This tendency arises from the sensible assumption that academicians of this kind are more closely concerned than anyone else with the selection of students for graduate study. However, the Advanced Tests are predictors of success in graduate study because they are measures of achievement in undergraduate study. Therefore, it is reasonable to have some professors who teach at the undergraduate level represented on the committees.

Disciplines differ in scope and homogeneity. Some are so broad they include several divisions that have themselves become virtually separate disciplines. Examples include zoology in biology, electrical engineering in engineering, American history in history, and Spanish American literature in Spanish. An important consideration in constituting a committee is adequate representation of the important divisions that exist in the discipline. Thus, the chemistry committee includes specialists in analytical, inorganic, organic, and physical chemistry, and the physics committee includes high-energy, nuclear, and solid-state physicists.

Another consideration in the appointment of committee members is geographical representation. In an increasingly mobile society in which a person may be born and reared in one region, do undergraduate work in a second, do graduate work in a third, and go on to become professor in a fourth, geographical representation is not an overriding consideration. However, this characteristic is a very easy one to ascertain, and an attempt is made to represent different regions of the country on each committee.

There is considerable concern, too, about fairly representing the interests of blacks, other minorities, and women on Advanced Test committees. Efforts to include at least one racial minority member and at least one woman on every committee have met with success.

The examiners are appointed by Educational Testing Service with the cooperation and assistance of a leading learned society (or societies) in the discipline of the test. The typical practice is for ETS to submit the names of several scholars under consideration for a [illegible] the executive secretary or president of the relevant [illegible] society officer is asked to comment on the scholars [illegible] consideration and to suggest additional scholars. The [illegible] for committee membership come to attention [illegible] channels. A scholar's writings in professional [illegible] at professional meetings may indicate special [illegible] interest in serving as an examiner. Sometimes [illegible] by current examiners who know what is in- [illegible] committee work and hence can recognize which of their colleagues may be especially suited for service on a committee. Sometimes a number of professors are invited to write questions for a test; those who are especially successful may later be appointed to a committee of examiners.

Each committee of examiners reviews and approves the specifications for the test for which it is responsible; writes, reviews, selects, revises, and approves questions for the test; and reviews and approves new editions of the test. Much of this work is carried out by mail, but typically each committee meets once for each new edition of a test that is developed. Thus, 12 committees meet annually, and 8 meet biennially.

Several characteristics of the Advanced Tests must remain fixed and can be changed only gradually. These stable characteristics reflect the fact that the GRE Advanced Tests are part of an ongoing testing program. For reasons of fairness, several editions of each Advanced Test must be available for administration each year. To be useful, these different editions must yield scores on the same scale. For the Advanced Tests, this means that each new edition must contain some questions from prior editions. It also means that the test specifications cannot change abruptly but must evolve over a period of time. Since all 20 Advanced Tests are given in the same 170-minute period in the same rooms, all tests must conform to the same administration mode. Finally, for reasons of effective measurement—that is, to provide measures of high validity and reliability with economy of time, money, and effort—the test questions must be of the multiple-choice type. Within this framework, the committees of examiners are free to exercise their judgment and creative skills in assessing the competencies of the examinees.

Content Specifications

When a new test is being introduced, the first task is to determine the test's future content and to set specifications. A number of problems must be faced and solutions devised. For example, committee members must grapple with such questions as: What are the major subareas of study within the field? Which are most important? With which subareas are enough students familiar to make the topics in those areas a reasonable focus of measurement? How can balance (knowledge and application, for example) among subareas and among skill dimensions be attained? If there are professional and academic tracks in a field of graduate study, which should be dealt with? Or can they be reconciled and both included? Is there a core of material basic to the study of the discipline or are specialized subfields relatively independent? Often no answer that serves the purposes of testing is fully satisfactory, but a consensus must be reached to provide a framework for future test development.

The first step in developing a new test edition after a test has been introduced is a review of test specifications. For many of the tests, two aspects of the specifications—subject matter to be covered and abilities to be measured—blend into a single dimension. By the time the examiners have satisfied themselves that a question requires the demonstration of a capability they believe to be a significant one for a graduate student in their discipline to possess, it is probably superfluous to inquire further into what name might be appropriately attached to the ability needed to answer the question. Hence, for many tests the specifications are set purely in terms of the subject areas of the discipline, with indications as to the number of questions to be included in each category of that content. For a few tests, the content and abilities are treated as two dimensions. Some attention is given to ascertaining, for example, if a question requires recall of information, application of information to the solution of a new problem, or analysis of a given body of information.

At each meeting of the committee, the test specifications are likely to be reviewed. The specifications agreed upon will guide test development until the next revision in the specifications. The practical implication of this procedure is that the specifications agreed upon at the first meeting are likely to guide development of the test prepared at a second meeting. Then the specifications agreed upon at the second meeting will guide development of the test prepared at a third meeting, and so on. The review of specifications focuses on such questions as: Has the field or the undergraduate curriculum changed in significant ways that will affect student knowledge? The newest trends in a field may not, of course, have yet had any effect on most curricula. Thus, the committee must consider the experience of the majority of students rather than the activities of a vanguard in the discipline.

One way of determining the appropriate content of the test is for committee members who have direct experience with students and with the teaching profession in the field to pool their knowledge of students' common experiences. Another method, particularly in cases where differences of opinion may exist, is to obtain information directly from the students taking the test. Periodically, students taking the Advanced Tests are asked to answer questions about their undergraduate background as well as their educational level (senior, graduate, etc.) and goal (master's, Ph.D., etc.). Most questions elicit information on specific courses taken or areas of concentration. (The answers to such questions for most of the Advanced Tests at 1970–71 test administrations are presented in Appendix II.) On occasion, more extensive questionnaires are distributed to examinees at test centers following the test; the examinees answer those questionnaires at home and mail their responses to ETS.

Whenever department heads or faculty members review a confidential inspection copy of a test, they are asked to complete a test evaluation form, expressing their judgments of the appropriateness of the test and of specific test questions for particular purposes. In the late 1960s and early 1970s, panels of faculty members at a number of institutions were systematically identified and asked to complete more detailed test evaluation forms. Several professors devoted many hours to these test reviews, and their collective judgments on test and question appropriateness proved valuable in determining test content.

Consultants may join the committees of examiners at their meetings occasionally. Often the consultant is an officer in the relevant professional association. Again, the discussion with the consultant is likely to focus on the appropriateness of the test content. From time to time, a panel of educators in a field may be convened to evaluate the test and make recommendations, or the faculty in an undergraduate field may be surveyed for reactions to test content.

Table 10: Statistical Characteristics of the Advanced Test Total Scores[1]

Advanced Test	Number of Examinees in Groups on which the Reported Scaled Scores Are Based	Reported Scaled Scores					Number of Examinees in Samples on which the Remaining Data Are Based	Difficulty	Reliability		Speededness	
		Mean	Standard Deviation	Highest Earned	Highest Possible	Lowest Possible		Average Percentage of Examinees Answering Questions Correctly	Reliability of Total Score[3]	Standard Error of Measurement of Total Scaled Scores[4]	Percentage of Examinees Completing ¾ of test	Percentage of Examinees Completing Entire test[5]
Biology	4,001	640	110	980	990 (1,060)	260	1,300	55%	.93	28	100%	90%
Chemistry	930	652	109	990 (1,010)[2]	990 (1,140)	440	930	39	.93	29	99	30
Computer Science	485	637	108	870	920	390	485	55	.93	28	100	54
Economics	600	615	114	960	990 (1,020)	400	600	43	.95	25	96	70
Education	2,164	458	91	700	810	220	2,160	49	.94	22	99	39
Engineering	1,310	614	109	910	990 (1,010)	320	1,310	50	.94	27	98	58
French	190	519	88	770	810	290	190	53	.96	19	100	62
Geography	259	469	91	690	850	210	255	50	.93	24	100	84
Geology	765	590	94	850	910	300	765	56	.94	22	99	75
German	320	527	100	760	810	290	320	57	.96	19	100	73
History	865	520	81	760	870	330	865	45	.94	20	99	89
Literature in English	1,378	548	101	800	810	250	1,375	60	.97	18	100	82
Mathematics	993	707	143	990 (1,060)	990 (1,060)	420	990	52	.93	38	100	68
Music	582	508	96	760	820	270	580	52	.96	19	99	64
Philosophy	225	660	118	960	990 (1,070)	380	225	48	.94	29	97	82
Physics	965	657	136	990 (1,090)	990 (1,210)	370	965	42	.89	45	99	32
Political Science	625	491	84	680	850	250	625	50	.92	24	99	91
Psychology	3,348	550	90	810	940	270	1,675	51	.93	25	100	58
Sociology	429	484	119	780	990 (1,000)	210	425	44	.94	29	99	81
Spanish	227	550	106	820	910	290	225	53	.95	23	100	92

[1] Values for editions introduced in 1976 (1974 for Geography and German).

[2] The numbers in parentheses represent scaled scores that would have been earned if the scale extended beyond 990. Scaled scores higher than 990 are reported as 990.

[3] Estimated by the Kuder-Richardson formula (20) adapted for use with formula scoring.

[4] This is the standard deviation of the scores an examinee would earn if taking the test repeatedly, assuming that factors such as fatigue were eliminable.

[5] As is noted on page 29, the percentage finishing the entire test may be misleading since the number reaching the last question may differ significantly from the number reaching the next-to-last question.

Statistical Specifications and Characteristics

The statistical specifications call for the Advanced Tests to be of middle difficulty, maximum reliability, and minimum speededness within the time constraints. Data on the difficulty, reliability, and speededness of the editions of each of the 20 tests introduced in 1976 (1974 for Geography and German) are given in Table 10. The data are obtained for each test edition introduced.

For maximum effectiveness in guiding admission decisions, the difficulties of questions should be such that about half the students who are at the dividing line between admission and rejection answer the questions correctly. If there were only one graduate school and the ability level of students who just qualified for admission to that graduate school could be established, then the test questions could be pitched at the ideal level of difficulty for making admission decisions at that school. Since there are hundreds of graduate schools, a reasonable alternative is to construct tests containing questions with a range of difficulties, but with an average question of such difficulty that half the examinees who respond to the question get it right.

In actuality, the Advanced Tests tend to be more difficult than the specifications recommended. For only 13 of the test editions in Table 10 is the average percentage of examinees answering questions correctly 50 percent or greater. One could infer that, for a five-choice question the answer to which is known by half the examinees and is guessed by the other half, the most likely percentage choosing the correct answer would be 60 percent. This is used as a reasonable standared of middle difficulty. Only one test is easy enough that the average percentage of examinees correctly answering the questions is 60 percent. The difficult questions are especially effective in distinguishing among higher-scoring examinees, whereas easy questions are especially effective in distinguishing among lower-scoring examinees.

If tests scores are to have value, they must possess a high degree of reliability. Reliability coefficients can range from 0.00 to 1.00, and meaning can be attached to them in several different ways. The method of determining reliability is discussed in Chapter 5. For the purposes of discussion here, reliability will be considered a statistical indicator of the tendency of a test to measure consistently from one time to another.

The reliabilities of the Advanced Tests almost always exceed .90. Such high reliabilities are desirable because the decisions being made partly on the basis of the tests are significant ones with considerable impact on people's lives. All the reliabilities in Table 10 equal or exceed .89, with six equal to or greater than .95. The standard errors of measurement range from 18 to 45 scaled score points, with a median of 25. Only two tests have standard errors of measurement greater than 29; they are the Advanced Mathematics and Physics Tests, which have relatively few questions. The Advanced Mathematics Test has 66 questions and the Advanced Physics Test 90. The nature of these fields, however, requires quite time-consuming questions. There is a reluctance to increase the number of questions because this might shift the measurement emphasis away from problem-solving ability toward the recall of facts. In January 1978 the committee of examiners for the Advanced Physics Test decided to raise the number of questions in the test to 100 to increase the reliability and decrease the standard error of measurement.

The Advanced Tests are intended to be of such a length that most examinees will have time to consider most, if not all, of the questions. Two speededness indicators are presented in Table 10; the percentage of students completing three-fourths of the test and the percentage answering the last question. Of the two speededness indicators given in Table 10, the percentage completing three-fourths of the test is the more reliable indicator. The percentage completing the test depends entirely on the number answering the last question. Often there is quite a large difference between the percentage reaching the next-to-last question and the percentage reaching the last question. One set of standards sometimes taken to indicate that a test is a power test and lacks any significant speed factor is that virtually all examinees reach three-fourths of the questions and 80 percent reach the last question. If "virtually all" is defined as 99 percent to 100 percent, then only three tests did not meet this standard. Although the tests are clearly designed to be power rather than speed tests, there is a strong desire to keep the number of questions very close to the level that tends to produce indications of speededness. The reason is that reliability depends on the number of questions included in the test. Observations show that examinees are quite variable in the rapidity with which they work. If the number of questions were reduced to allow at least 80 percent to complete every test, much useful information would be sacrificed, and the test reliabilities would suffer accordingly.

In the assembly of most Advanced Test editions, computed statistical information is available only for the equating questions and a few other questions. If the new edition is being equated through one prior edition, statistical information will be available for about 20 percent of the questions. If it is being equated through two prior editions, as is often done, information will be available for about 40 percent of the questions. New questions are not pretested; thus their statistical characteristics must be estimated.

Pretesting of questions for the Advanced Tests was introduced at one time but was later discontinued. Pretesting does permit construction of tests whose actual characteristics will more closely meet specifications than otherwise. However, it has proved possible to assemble Advanced Tests with fully adequate statistical characteristics without pretesting new questions. (See the discussion of pretesting in Chapter 2.)

Two or three subscores are reported for nine Advanced Tests. The subscores are intended to provide information useful to

Table 11: Statistical Characteristics of the Advanced Test Subscores*

Advanced Test	Subscores	Number of Examinees in Groups on which the Reported Scaled Subscores Are Based	Reported Scaled Subscores		Number of Examinees in Samples on which the Remaining Data Are Based	Difficulty	Reliability	
			Mean	Standard Deviation		Average Percentage of Examinees Answering Questions Correctly	Reliability of Subscore	Standard Error of Measurement of Subscore (Scaled)
Biology	Cellular & Subcellular	4,001	64	11	1,300	53%	.88	3.8
	Organismal		64	11		53	.82	4.7
	Population		64	11		59	.85	4.2
Engineering	Engineering	1,310	61	11	1,310	41	.86	4.1
	Mathematics Usage		61	11		60	.92	3.0
French	Interpretive Reading Skills	190	52	9	190	63	.93	2.4
	Literature & Civilization		52	9		43	.92	2.5
Geography	Human	255	47	9	255	53	.89	3.0
	Physical		47	9		45	.88	3.1
Geology	Stratigraphy, Paleontology, and Geomorphology	765	59	9	765	58	.84	3.7
	Structural Geology and Geophysics		59	9		56	.87	3.4
	Mineralogy, Petrology, and Geochemistry		59	9		53	.88	3.3
History	European	865	52	8	865	45	.92	2.3
	American		52	8		44	.85	3.1
Music	Theory	582	51	10	580	58	.90	3.1
	History		51	10		48	.94	2.3
Psychology	Experimental	3,348	55	9	1,675	47	.87	3.3
	Social		55	9		56	.85	3.5
Spanish	Interpretive Reading Skills	227	55	11	225	54	.88	3.8
	Peninsular Topics		55	11		51	.89	3.7
	Spanish-American Topics		55	11		46	.85	4.2

*Values for editions introduced in 1976 (1974 for Geography).

students in assessing their strengths and weaknesses and useful to institutions in guiding and placing students. Subscores are reported if the committee of examiners for a test identifies subscores judged to serve these purposes and if the reliabilities of the subscores exceed .80. The number of subscores represents a compromise between the desire to report many subscores to more fully describe the student's achievement and the desire to report subscores of sufficient reliability to be used with confidence. The reliability of a subscore depends in part on the number of questions contributing to it. As the number of subscores increases, the number of questions contributing to a subscore and hence the reliability of that subscore tend to decrease. Since subscores tend to have lower reliabilities than total scores, they are more appropriately used for placement than admission decisions. Data on subscore performance, the average difficulty of questions contributing to each subscore, and the reliability of subscores are given in Table 11. All but two of the subscores have reliabilities at or above .85.

Table 12 shows the correlations of the subscores with each other and with the total scores. The correlations (between .52 and .82) indicate that the subscores are sufficiently independent to be useful in providing more *specific* information about the student's relative achievement in the subdivisions of the field. The correlations between subscores and total scores are spuriously high because the questions in each subscore make up a substantial portion, sometimes more than half, of the questions contributing to the total score.

Most students who take Advanced Tests also take the GRE Aptitude Test. Correlations between verbal and quantitative ability scores on the Aptitude Test and Advanced Test scores are shown in Table 13. These data are from 1967–68, the last time correlations between Advanced Test and Aptitude Test scores were calculated. The median correlation between Advanced Test score and verbal ability score was .63, and the median correlation between Advanced Test score and quantitative ability score was .52. These correlations suggest that the Advanced Tests measure domains substantially independent of the verbal and quantitative ability measures of the Aptitude Test.[*]

---

[*] Correlations between Advanced Test scores and the restructured Aptitude Test scores, including the analytical ability score, were not available at the time of this manual's publication.

Table 12: Correlations among Scores on the Advanced Tests for which Subscores Are Reported

Advanced Test	Score	Total Score	Subscore		
			(1)	(2)	(3)
BIOLOGY:	Total Score	1.00	.84 to .87	.89 to .92	.85 to .87
	(1) Cellular and Subcellular Biology	.84 to .87	1.00	.64 to .68	.52 to .58
	(2) Organismal Biology	.89 to .92	.64 to .68	1.00	.70 to .73
	(3) Population Biology	.85 to .87	.52 to .58	.70 to .73	1.00
ENGINEERING:	Total Score	1.00	.90 to .91	.91 to .92	
	(1) Engineering	.90 to .91	1.00	.65 to .67	
	(2) Mathematics Usage	.91 to .92	.65 to .67	1.00	
FRENCH:	Total Score	1.00	.93 to .94	.92 to .93	
	(1) Interpretive Reading Skills	.93 to .94	1.00	.72 to .74	
	(2) Literature and Civilization	.92 to .93	.72 to .74	1.00	
GEOGRAPHY:	Total Score	1.00	.92 to .96	.85 to .88	
	(1) Human Geography	.92 to .96	1.00	.63 to .68	
	(2) Physical Geography	.85 to .88	.63 to .68	1.00	
GEOLOGY:	Total Score	1.00	.89	.90 to .91	.86 to .89
	(1) Stratigraphy, Paleontology, Geomorphology	.89	1.00	.71 to .73	.63 to .70
	(2) Structural Geology, Geophysics	.90 to .91	.71 to .73	1.00	.68 to .70
	(3) Mineralogy, Petrology, Geochemistry	.86 to .89	.63 to .70	.68 to .70	1.00
HISTORY:	Total Score	1.00	.95 to .97	.89 to .90	
	(1) European History	.95 to .97	1.00	.72 to .76	
	(2) American History	.89 to .90	.72 to .76	1.00	
MUSIC:	Total Score	1.00	.92 to .93	.96 to .97	
	(1) Theory of Music	.92 to .93	1.00	.77 to .82	
	(2) History of Music	.96 to .97	.77 to .82	1.00	
PSYCHOLOGY:	Total Score	1.00	.90 to .93	.90 to .92	
	(1) Experimental Psychology	.90 to .93	1.00	.67 to .73	
	(2) Social Psychology	.90 to .92	.67 to .73	1.00	
SPANISH:	Total Score	1.00	.91 to .93	.87 to .88	.80 to .85
	(1) Interpretive Reading Skills	.91 to .93	1.00	.69 to .70	.63 to .72
	(2) Peninsular Topics	.87 to .88	.69 to .70	1.00	.63 to .65
	(3) Spanish-American Topics	.80 to .85	.63 to .72	.63 to .65	1.00

Table 13: Correlations of Advanced Test Scores with Aptitude Test Scores, 1967-68

Advanced Test	Number of Examinees	Correlations Between:	
		Advanced Test and Verbal Ability Scores	Advanced Test and Quantitative Ability Scores
Biology	4,696	.66	.60
Chemistry	2,486	.50	.58
Computer Science	(No test in 1967-68)		
Economics	1,930	.66	.65
Education	2,746	.71	.46
Engineering	4,259	.49	.61
French	1,292	.69	.41
Geography	306	.53	.54
Geology	575	.51	.50
German	(No test in 1967-68)		
History	4,919	.61	.40
Literature in English	6,276	.75	.40
Mathematics	3,279	.55	.65
Music	647	.60	.54
Philosophy	793	.69	.50
Physics	2,190	.46	.55
Political Science	2,745	.65	.50
Psychology	5,643	.67	.50
Sociology	2,151	.78	.67
Spanish	770	.39	.18

The six Advanced Tests with the highest correlations with verbal ability scores were, in order, Sociology, Literature in English, Education, French, Philosophy, and Psychology. The six Advanced Tests with the highest correlations with quantitative ability scores were, in order, Sociology, Economics, Mathematics, Engineering, Biology, and Chemistry. The significance of these correlations is discussed in Chapter 6 in relation to the construct validity of the Aptitude Test.

Information Unique to Each Advanced Test

Several important categories of information about each of the Advanced Tests are treated in Appendix II, which provides for each Advanced Test, where appropriate and available, the following:

1. A description of the test's content, specifications, and specific problems associated with determining the test content.
2. Responses of students to questions about undergraduate background in the discipline.
3. Reports of validity studies or other studies involving the test.

Table 14 provides a historical perspective on each Advanced Test offering by listing tests available in 1956–57, in 1966–67, and in 1976–77, with indications of changes that took place in the decades between.

Reference

Graduate Record Examinations Board and Council of Graduate Schools in the United States. *Graduate Programs and Admissions Manual, 1976–77*. Princeton, N.J.: Educational Testing Service, 1976.

Table 14: Advanced Tests Available

16 Tests in 1956-57	Changes in 1957-1966	21 Tests in 1966-67	Changes in 1967-1976	20 Tests in 1976-77
Biology Chemistry Economics Education Engineering French Geology Government History Literature Mathematics Philosophy Physics Psychology Sociology Spanish	Business, Geography, Music, Physical Education, and Speech were added.	Biology Business Chemistry Economics Education Engineering French Geography Geology Government History Literature Mathematics Music Philosophy Physical Education Physics Psychology Sociology Spanish Speech	Business, Physical Education, and Speech were dropped; Computer Science and German were added; Anthropology was added, then dropped. The name of Government was changed to Political Science, and the name of Literature was changed to Literature in English.	Biology Chemistry Computer Science Economics Education Engineering French Geography Geology German History Literature in English Mathematics Music Philosophy Physics Political Science Psychology Sociology Spanish

Chapter 5

STATISTICAL METHODS AND ANALYSES OF THE GRADUATE RECORD EXAMINATIONS

The psychometric problems of the Graduate Record Examinations Program are similar to those found in any testing program with the following characteristics: (a) the services offered are based on a battery of tests rather than on a single test; (b) the tests are administered more than once a year; (c) they are administered over an extended period of years; and (d) the complex nature of the services includes providing information used in making decisions that have a long-term effect on individuals and are important for institutions.

In a testing program in which only one test and one administration are involved, the score scale may be defined in any convenient and arbitrary fashion. For example, the score scale for grading a final examination prepared for evaluating one class of students is selected by the person who does the evaluating and adequately serves the dual purposes of ranking the students in that class and establishing an acceptable level of performance. No further use of the scale is anticipated.

However, when more than one test or score is involved, it may be desirable to introduce some kind of linkage that binds each test scale to the others in a manner that will facilitate score interpretation. Up until 1977, the Aptitude Test provided two basic measures of academic potential that have been widely used in admissions decisions. It is obvious that reporting the verbal and quantitative ability scores on a single scale was a necessary convenience. It is equally obvious that the new analytical ability measure, introduced as part of the Aptitude Test in the fall of 1977, be reported on that same scale. In the case of the Advanced Tests, however, it is not so obvious, but the early years of experience with the development of the Graduate Record Examinations showed that there were decided advantages in having a scale structure that would reflect the relative ability levels of students who elect the various major fields. This is the kind of scaled-score system currently used in the GRE Program.

A testing program that involves multiple administrations within a relatively short period of time must provide alternate forms of each test and comparability of scores across test forms. Each alternate form must satisfy to a high degree all the requirements of parallelism in both content and statistical characteristics with the form it is to replace. Because alternate forms will inevitably differ somewhat in difficulty level, some statistical adjustments must be made in the score conversion to make the scores reported on the two forms directly comparable. This adjustment is accomplished through the statistical procedure of equating.

When a test is administered frequently over an extended period of time, it becomes necessary to insure the stability of the scale so that the experience gained in interpreting scores earned in previous years can be applied to interpreting scores earned in the current year. This requires that test forms be indeed parallel so as to minimize the statistical errors of equating. At the same time, however, there is a psychometric need to allow for revisions in content that reflect advancements in knowledge and changing curricular emphasis within a field. The conflict between the statistical requirements for scale stability and the psychometric requirements for test vitality is resolved by permitting gradual content revision combined with tight statistical control of the scale.

Although scaling and equating procedures provide for consistency in reporting scores, the interpretation of those scores must be enhanced by information on the ranking of each score among all other scores for some meaningful group. Therefore, normative information is provided along with other interpretive data to help students and graduate school representatives compare the performance of an individual with that of others. The *Guide to the Use of the Graduate Record Examinations* provides three sets of normative or interpretive tables. The first set provides percentile ranks for selected scaled scores for the total GRE population taking a particular test within a recent three-year period. The second set provides the same kind of information for recent GRE National Program seniors and nonenrolled college graduates (the typical applicants for admission to graduate schools, approximately 60 percent of the total GRE population) who have taken the Aptitude Test and may have taken an Advanced Test. The third set, which is based on this same group, provides Aptitude Test score distributions based on the classification of the examinees by intended graduate major field. (As data are accumulated for the restructured Aptitude Test, first administered in October 1977, they will be presented in succeeding issues of the *Guide*.)

Development of the GRE Scaled-Score System

The scaled-score system used in the GRE Program defines a scaled score of 500 as the mean of the score distribution for the particular standardization or reference group on which the scale is based and 100 as the standard deviation of that distribution. Scores are reported as three-digit scores ending in zero and having a maximum permitted range of 200 to 900 for the Aptitude Test and 200 to 990 for the Advanced Tests. Prior to 1952 each test was scaled independently of the others by setting the mean of the group that took the test equal to 500 and the standard deviation equal to 100. Shortly before 1952, the Advanced Tests were extensively revised and the allotted testing time was extended from one hour and forty-five minutes to three hours. As a result of these changes in the tests and changes in the GRE population, a decision was made to rescale the tests and to recognize the advantages of changing the type of scaled-score system used.

Students majoring in different fields generally exhibit different levels and ranges of aptitude development. In the redesign of the scaled-score system, these differences were to be taken into account and incorporated in the scales for the individual Advanced Tests. The data for the rescaling were collected in the spring of 1952 and consisted of scores earned by 2,095 graduating seniors in 11 colleges. This group was considered at that time to be reasonably representative of the GRE senior population, and, therefore, a scale system based on their performance would have normative properties useful in the interpretation of scores obtained by other groups in subsequent administrations.

Each student in the 1952 scaling group took the Aptitude Test and also the Advanced Test appropriate to his or her major field. The scale for each score (verbal ability and quantitative ability) on the Aptitude Test was established by setting the total-group mean equal to 500 and the standard deviation equal to 100. This process, which preserved the rank order of the students, resulted in a linear transformation for converting raw scores to scaled scores.

For the scaling of the Advanced Tests, a more complex statistical procedure was employed. For each Advanced Test subgroup, regression coefficients were determined for predicting Advanced Test scores from the verbal and quantitative ability scores on the Aptitude Test. Estimations were then made of the raw-score mean and standard deviation of the entire standardization group on that Advanced Test. The *estimated* mean was then set equal to 500 and the *estimated* standard deviation equal to 100. The equations for estimating the mean and standard deviation were developed by Ledyard R Tucker and are reprinted below from the article written by Schultz and Angoff (1956) describing the 1952 scaling experiment.

Notation:

v, q = scaled scores on the verbal and quantitative parts of the Aptitude Test
x = raw scores on the Advanced Test
t = entire standardization group (2,095 seniors)
s = Advanced Test subgroup
$\tilde{M}, \tilde{\sigma}^2$ = estimated mean and variance
M, σ^2 = obtained mean and variance
b = regression coefficient
C_{vq} = covariance between verbal and quantitative

Table 15: Scaled-Score Means and Standard Deviations of the 1952 Standardization Group

Advanced Test Subgroup	N	Verbal Ability		Quantitative Ability		Advanced Test	
		Mean	S.D.	Mean	S.D.	Mean	S.D.
Biology	209	486	94	499	87	495	96
Chemistry	180	507	96	562	99	530	101
Economics	239	476	89	516	99	494	97
Education	180	438	86	434	83	446	93
Engineering	151	454	92	570	86	497	98
French	32	520	92	453	72	533	92
Geology	35	473	92	500	86	488	97
German*	10	543	69	495	79		
History	181	517	93	468	80	506	97
Literature	239	564	98	463	86	548	99
Mathematics	81	508	93	587	90	542	97
Philosophy	31	563	96	521	90	549	97
Physics	49	531	106	633	79	546	101
Political Science	146	498	93	485	89	496	97
Psychology	171	527	94	495	86	512	96
Sociology	127	482	93	447	77	474	96
Spanish	34	529	102	451	83	520	99
Entire Scaling Group	2,095	500	100	500	100		

*The German subgroup was too small for use in establishing the scale for this Advanced Test. A revised version of the test was scaled in 1969 in the Undergraduate Program.

[The information in the above table was taken from Schultz and Angoff (1956).]

The estimated mean and the estimated variance for the 2,095 seniors in the standardization group are given by the equations:

$$\tilde{M}_{x_t} = M_{x_s} + b_{xv.q_s}(M_{v_t} - M_{v_s}) + b_{xq.v_s}(M_{q_t} - M_{q_s}),$$

$$\tilde{\sigma}^2_{x_t} = \sigma^2_{x_s} + b^2_{xv.q_s}(\sigma^2_{v_t} - \sigma^2_{v_s}) + b^2_{xq.v_s}(\sigma^2_{q_t} - \sigma^2_{q_s})$$
$$+ 2b_{xv.q_s}\, b_{xq.v_s}(C_{vq_t} - C_{vq_s}).$$

The linear transformation to obtain scaled scores (y) from raw scores (x) is defined by the equation:

$$y = A\,x + B,$$

where

$$A = \frac{100}{\tilde{\sigma}_{x_t}} \quad \text{and} \quad B = 500 - A\,\tilde{M}_{x_t}$$

As can be seen in Table 15, the Advanced Test mean for the subgroup that actually took that test reflects the aptitude level of that group. This kind of difference in Aptitude Test performance was taken into account in establishing the scales for the Advanced Tests.

As each new test was added to the GRE battery, it was placed on the GRE scale in a similar manner. The scaling sample consisted of the first group of seniors who took the new test along with the Aptitude Test. The scale was established by using the relationships among the three scores, thus reflecting the aptitude level and range of the group tested. The following tests were introduced after the 1952 scaling administration:

Anthropology	(scaled in 1968, withdrawn in 1971)
Business	(scaled in 1964, withdrawn in 1970)
Computer Science	(scaled in 1976)
Geography	(scaled in 1966)
German	(scaled in 1969 in the Undergraduate Program as a Field Test; GRE counterpart equated to the UP test in 1970)
Music	(scaled in 1964 for the GRE Institutional Program, added to the GRE National Program in 1965)
Physical Education	(available in 1962, extensively revised and rescaled in 1965, withdrawn in 1970)
Speech	(scaled in 1953, withdrawn in 1970)

For each of these tests, the scaling process involved using the verbal and quantitative ability scores of the group taking the test to estimate the Advanced Test performance of the *original 1952 standardization group* and setting the *estimated* mean equal to 500 and the *estimated* standard deviation equal to 100. A significant difference between the scaling of these tests and the scaling of those included in the 1952 scaling experiment lies in the fact that the relationships among the verbal and quantitative ability and Advanced Test scores are based on more recent populations. The possible effect this might have on interpretation of the scores is discussed in the section dealing with the rescaling study of 1967–68 (page 38.)

Scaling of the Analytical Ability Measure

The introduction of the new analytical ability measure in the GRE Aptitude Test in October 1977 presented a new scaling problem. If the analytical measure had been introduced with the verbal and quantitative measures in the original 1952 scaling administration, the analytical ability scores would have been put on scale by setting the mean of the standardization group equal to 500 and the standard deviation equal to 100, as was done with the verbal and quantitative ability scores. Obviously, this procedure was not possible in 1977.

The correlations of the analytical ability measure with the verbal and quantitative ability measures are nearly equal and are rather high: approximately .76 for verbal and .74 for quantitative. The method selected for scaling the analytical ability score consisted of averaging the verbal and quantitative ability score means and variances using β-weights, as shown in the following formulas:

Estimate of the analytical score mean:

$$\tilde{M}_a = \frac{\beta_{av.q} M_v + \beta_{aq.v} M_q}{\beta_{av.q} + \beta_{aq.v}},$$

and of the variance:

$$\tilde{\sigma}_a^2 = \frac{(\beta_{av.q}\sigma_v + \beta_{aq.v}\sigma_q)^2}{(\beta_{av.q} + \beta_{aq.v})^2},$$

where

$$\beta_{av.q} = \frac{r_{av} - r_{aq} r_{vq}}{1 - r^2_{vq}},$$

$$\beta_{aq.v} = \frac{r_{aq} - r_{av} r_{vq}}{1 - r^2_{vq}},$$

and the means and variances are in scaled-score units. The resulting means and standard deviations for the October 1977 scaling administration are: 503 and 126 for verbal ability, 525 and 133 for quantitative ability, and 513 and 129 for analytical ability.

Score Equating and Related Concerns

The purpose of equating is to permit introduction of new forms of a test to replace old forms without losing comparability of reported scores and long-term continuity of the established score scale. Among the conditions for sound equating are four of particular importance: the new form must be parallel to the old form; the equating samples must be adequate in size and must represent the population for which the test was designed; the conditions of test administration must be carefully controlled; and the equating method must be appropriate for the particular equating experiment. In practice, however, compromises are sometimes necessary to accommodate other considerations governing the policies of the testing program.

A rigid interpretation of the conditions of parallelism would result in production of a test built with the same content specifications, the same number of test items,[*] the same level and spread of difficulty, and the same reliability as the old form. Even with this objective governing test construction, the requirements for parallelism can never be precisely met, but the deviations are small and the statistical adjustments of the equating are adequate. Advanced Test committees of examiners may from time to time reassess the content specifications and possibly make changes that reflect changes in curricula and in populations. An examination of the content specifications of the old form may indicate that the difficulty level is no longer appropriate for current groups or that the test length should be changed. Changes of this kind, however, are purposely introduced gradually to avoid a serious effect on the comparability of scores across forms.

In the planning of an equating experiment, every reasonable effort is made to establish samples that are large and representative. Practical circumstances do occasionally necessitate compromise. Sample size is limited by the nature of the test and the number of examinees tested in one administration. For example, geography is a very small-volume field, and it is unlikely that more than 150 examinees will take the Advanced Geography Test at one time, whereas the volume for the Aptitude Test will exceed 50,000 examinees. In the case of Advanced Geography, the small sample size is balanced to some degree by the fact that the sample represents the population. In all cases, successful GRE equating is facilitated by the standardization and tight control of the test administrations.

The choice of equating method is dictated in part by sampling possibilities and in part by test construction. The Aptitude Test is perhaps the most important test in the GRE battery and the most widely used. For security reasons, it is constructed so that each test item appears in only one form. This fact, combined with the availability of large samples, permits the use of an equating method that should not be used for the Advanced Tests.

A review of the most fundamental method of equating is a useful preface to discussion of the methods used for equating the GRE Aptitude Test and Advanced Tests. In this method, the old form and the new form of a test are administered to the same group of examinees, and the assumption is made that performance on the second form is in no way affected by the fact that the examinees have previously taken the first form. The scale of measurement of the second form is then transformed in such a way that the frequency distribution of the transformed scores will be statistically equivalent to that of the raw scores on the first form. This can be accomplished through a linear transformation that sets the raw-score mean and standard deviation of the new form equal to the corresponding raw-score statistics of the old form:

Let

X = raw score on the new form,
Y = raw score on the old form,
σ_x, M_x = raw-score mean and standard deviation for the new form,
σ_y, M_y = raw-score mean and standard deviation for the old form,
a_x, b_x = conversion parameters for transforming raw scores on X to the scale of Y.

Then, $M_y = a_x M_x + b_x,$
and $\sigma_y = a_x \sigma_x.$

$$\left.\begin{aligned} a_x &= \sigma_y/\sigma_x \\ b_x &= M_y - a_x M_x. \end{aligned}\right\} \text{Conversion Parameters}$$

These parameters convert the X scores to scores on the Y raw-score scale. If the old form has already been put on a scale for reporting, it has conversion parameters A_y and B_y. Thus, the scores on the new form can be put on the same scale for reporting by using the transformation:

$$\text{Scaled Score} = A_x X + B_x,$$

where

$$A_x = A_y a_x \quad \text{and} \quad B_x = A_y b_x + B_y.$$

This fundamental method has a serious weakness in that the assumption on which it is based is not valid. The effects of practice and/or fatigue are disturbing factors that make the method unacceptable in almost all circumstances. However, the idea of having the group taking the new form equivalent to the group taking the old form is sound and forms the basis for other equating methods. The methods actually used currently in the GRE Program are discussed in the next sections.

---

[*] All GRE test questions are objective. Since the conventional psychometric term for an objective test question is *item*, we will use item throughout the rest of this chapter.

Aptitude Test Equating. Because the examinee volume for the Aptitude Test is so large, the problem of getting two equating samples (one for the new form and one for the old form) that can be considered equivalent is solved by making use of a test administration practice called "spiraling." The test books are packaged in spiraled order, alternating Form A (the old form) with Form B (the new form) in such a manner that at every testing center half the examinees take Form A and the other half Form B. Because the volume insures samples of more than 10,000 cases, the two groups are considered comparable and the random errors of sampling negligible. The assumption is then made that the scaled-score mean and standard deviation for the group taking the new form should be equal to the corresponding measures for the old form. The computation is the same as that described in the introduction to equating.

Under certain circumstances, spiraling cannot be used. For example, when the timing of the new form is different from that of the old form, the two forms cannot be administered together in the same testing room. This was the situation with the introduction of the restructured Aptitude Test in October 1977. The equating problem was solved by using an external equating subtest to establish a link between the old form and the new one. In the January 1977 administration, four different versions of a current form of the Aptitude Test were used to establish four old-form equating samples. Two versions included verbal equating subtests as Section IV (the pretest section) and two had quantitative equating subtests. In the October 1977 administration, each of the two new forms also had four versions with the same four equating subtests. Each equating subtest was used as common material to link the new verbal measure (or quantitative measure) to its old counterpart, thus providing two independent links between each new and old form. The equations used to establish these relationships are discussed in the following section on Advanced Test equating.

Advanced Test Equating. The relatively small samples available for equating the Advanced Tests make the spiraling technique undesirable because the random sampling error would be unacceptably high. Therefore, a different procedure for establishing equivalent samples is required. The new-form sample generally consists of all examinees tested in the current administration. The old-form sample is selected from groups tested in previous administrations and is matched, insofar as possible, to the expected performance level of the new-form group. Both forms of the test contain a common subset of items that represents the total test in content and statistical properties. Since the two forms of the test are parallel, it is reasonable to assume that the practice effect of taking the equating subset on the scores obtained on the total test is the same for both groups. The observed relationships between total score and equating subset score for the two groups are used to make a statistical *estimate* of the total-score mean and standard deviation of the combined groups on each of the two forms. Thus, we now have one sample (the new- and old-form samples combined into one sample), with *estimated* mean and standard deviation on the new form and *estimated* mean and standard deviation on the old form. These estimated values are then set equal to each other by applying the procedures described in the introduction to equating.

Two methods for estimating the means and standard deviations are used. The first, proposed by Ledyard Tucker and referred to as the Tucker equations, is appropriate when the new-form and the old-form samples are well matched. The equations for estimating are given below in the notation used in the introduction to this section, expanded to include sample identification and several additional terms.

Let

X = the score on the new form taken by group α,
Y = the score on the old form taken by group β,
V = the score on the equating subtest taken by the total group t, where $t = \alpha + \beta$,
$\tilde{M}_{x_t}$ = the estimated mean of the total group on Form X,
$\tilde{C}_{xx_t}$ = the estimated variance (square of the standard deviation of the total group on Form X).

Then $$\tilde{M}_{x_t} = M_{x\alpha} + \frac{C_{xv_\alpha}}{C_{vv\alpha}}\left(M_{v_t} - M_{v\alpha}\right).$$

and $$\tilde{C}_{xx_t} = C_{xx\alpha} + \left(\frac{C_{xv_\alpha}}{C_{vv_\alpha}}\right)^2\left(C_{vv_t} - C_{vv\alpha}\right).$$

These equations provide estimates of the total-group mean and variance on Test X. There is a parallel set of equations for Test Y based on the observed statistics of the β group. From this point on, the procedures are those described previously for equating means and standard deviations. The disadvantage of this method is that it is based on two conditions that are not always satisfied in practice: that the two forms are parallel and that the two samples are similar.

The Levine (1955) equations (also called the major axis equations) were developed for use in those situations not appropriate for the Tucker equations. In practice it is not always possible to select similar equating samples, and the Levine equations are preferred in this case. There are four sets of equations for estimating total-group means and variances when both total test and equating subtests are scored in the same way. (In the GRE Program scores are all computed by the formula: Rights − k × Wrongs.) Two sets are based on the condition of equal reliabilities of the two forms, and the other two on the condition of unequal reliabilities. Each of these pairs is further categorized by the location of the equating subtest: internal (included in the total score) or external (equating subtest separate from the total test). For most Advanced Test equating experiments, the reliabilities are assumed to be equal (same timing, same number of items, and parallel forms) and the equating subtest is part of the total test. For some experiments, the reliabilities are assumed to be unequal because a significant change has occurred. On rare occasions, the equating subtest may be administered as a separate test.

Using the same notational system that was used in the preceding explanations, we have four sets of equations for estimating total-group means and variances:

1. Equal reliabilities, equating subtest included in the total test:

$$\tilde{M}_{x_t} = M_{x_\alpha} + \frac{C_{xx_\alpha}}{C_{xv_\alpha}}\left(M_{v_t} - M_{v_\alpha}\right).$$

$$\tilde{C}_{xx_t} = C_{xx\alpha} + \left(\frac{C_{xx_\alpha}}{C_{xv_\alpha}}\right)^2\left(C_{vv_t} - C_{vv\alpha}\right).$$

For this case, and also for the three other cases, there is a parallel set of equations for Form Y based on the observed statistics of the β group. This set of equations is the one most frequently used in equating the GRE Advanced Tests.

2. Unequal reliabilities, equating subtest included in the total test:

$$\tilde{M}_{x_t} = M_{x_\alpha} + \frac{C_{xx_\alpha}}{C_{xv_\alpha}}\left(M_{v_t} - M_{v_\alpha}\right).$$

$$Q\tilde{C}_{xx_t} = \left(\frac{C_{xx_\alpha}}{C_{xv_\alpha}}\right)^2$$

The factor Q in the estimate of the total-group variance appears in both estimates of variance (for Test Y as well as for Test X) and consequently drops out in the computation of the conversion parameter a_x. Therefore, there is no need to compute it. These equations are used for equating Advanced Tests that have undergone some change. When the timing, for example, was decreased from 180 minutes to 170 minutes in 1972, this method was used for all the tests.

3. Unequal reliabilities, equating subtest external to the total test:

$$\tilde{M}_{x_t} = M_{x_\alpha} + \frac{C_{xx_\alpha} + C_{xv_\alpha}}{C_{vv_\alpha} + C_{xv_\alpha}}\left(M_{v_t} - M_{v_\alpha}\right).$$

$$Q\tilde{C}_{xx_t} = \left(\frac{C_{xx_\alpha} + C_{xv_\alpha}}{C_{vv_\alpha} + C_{xv_\alpha}}\right)^2.$$

As in the second case given above, the factor Q appears in both estimates of variance and consequently drops out in the computation of a_x, and there is no need to compute it. This method is used when the new form has no items in common with the old form and differs from it in a nontrivial way. This was the case with the equating of the verbal and quantitative parts of the first restructured Aptitude Test forms introduced in October 1977. It would be used with an Advanced Test under unusual circumstances, but this is not likely to happen.

4. Equal reliabilities, equating subtest external to the total test:

$$\tilde{M}_{x_t} = M_{x_\alpha} + \frac{C_{xx_\alpha} + C_{xv_\alpha}}{C_{vv\alpha} + C_{xv\alpha}}\left(M_{v_t} - M_{v_\alpha}\right).$$

$$\tilde{C}_{xx_t} = C_{xx_\alpha} + \left(\frac{C_{xx_\alpha} + C_{xv_\alpha}}{C_{vv_\alpha} + C_{xv_\alpha}}\right)^2\left(C_{vv_t} - C_{vv_\alpha}\right).$$

This fourth set of equations has not yet been used in the Graduate Record Examinations Program, but it may be used for Aptitude Test equating in the future. The strongest advantage of this method is that it permits the introduction of a new form that has no items in common with previous forms.

In a practical equating experiment, it is usually not possible to predict that the two equating samples will be well matched. Therefore, the common practice is to apply both the Tucker and the appropriate Levine methods and exercise the option of choosing the more appropriate method when the sample statistics become known. This is done by comparing the performance of the two samples on the common measure, the equating subset V:

$$\left|\frac{M_{v_\alpha} - M_{v_\beta}}{\sigma_{v_t}}\right| < 0.25$$

[Test of significance for the difference of the means]

$$0.80 < \frac{C_{vv_\alpha}}{C_{vv_\beta}} < 1.25$$

[Test of significance for the difference of the standard deviations]

If both these statements are true in a given equating experiment, then the two samples are sufficiently similar and the Tucker method is preferred. If either one is not true, then the samples are sufficiently different to make the Levine estimates the better choice.

Subscore Scaling

Subscore reporting was first used in the Graduate Record Examinations Program in 1965, when a revised version of the Advanced Speech Test was introduced. (The test was withdrawn in 1970.) Subscore reporting was next used with the Advanced Geography Test in 1969. In 1972, subscore reporting service was expanded and now includes nine Advanced Tests: Biology, Engineering, French, Geography, Geology, History, Music, Psychology, and Spanish. In each case, the subscores for the first form were scaled by setting the subscore mean equal to one-tenth the total-score scaled-score mean and the subscore standard deviation equal to one-tenth the total-score scaled-score standard deviation. Thus, the subscore scale is a two-digit scale directly related to the total-score scale and having a maximum permitted range of 20 to 99. For example, if an examinee has an Advanced Biology score of 600 and is a biology major, his or her performance on the three subscores is likely to be in the range of 56 to 64.

After the initial scaling of the subscores, some problems were encountered in using common-item equating (equating subtest) for the subscores of subsequent forms. Both the Tucker and the Levine methods require each new form to have items in common with the old form. For sound equating, each test or subtest must include at least 20 items in common with the old form. Therefore, a test with three subscores would require at least 60 items in common with the old form, and that amount of overlap is not acceptable for security reasons. A compromise solution to this problem is to scale the subscores of each new form through the total score, as was done in the initial scaling of the subscores. In the case of the Advanced Geography Test, the results of this kind of equating were compared with the common-item equating method in 1969, and the two methods were found to yield almost identical results. Since that experiment, the scaling procedure has been used to place the subscores of new forms on the GRE subscore scale for reporting.

Stability of the Scale

The preceding discussion of total-score equating explains how the form-to-form relationships are established. It implies a sequential operation extending over a long period of time in which each new form is equated to its immediate predecessor. If that procedure were followed in practice without modification, the stability of the scale would be in jeopardy and scale "wobble" could develop. This presents a serious problem in a program such as GRE, in which as many as five test forms may be used interchangeably in one academic year and more than ten forms in a five-year period. Comparability of scores across forms is such an essential part of the GRE score reporting service that considerable effort is expended in establishing and maintaining scale stability.

One solution to this problem is the use of a statistical technique called double part-score equating. In this procedure, each new form of a test is equated to *two* old forms instead of one, thus obtaining two conversion lines relating scores on the new form to the GRE scale. The final conversion line is the bisector of the angle formed by these two lines:

Let

$Y = A_1 X + B_1$ be the conversion line obtained by equating Test X to one old form,

and

$Y = A_2 X + B_2$ the conversion line obtained by equating Test X to a second old form.

Then, the bisector of these two lines is $Y = A X + B$, where

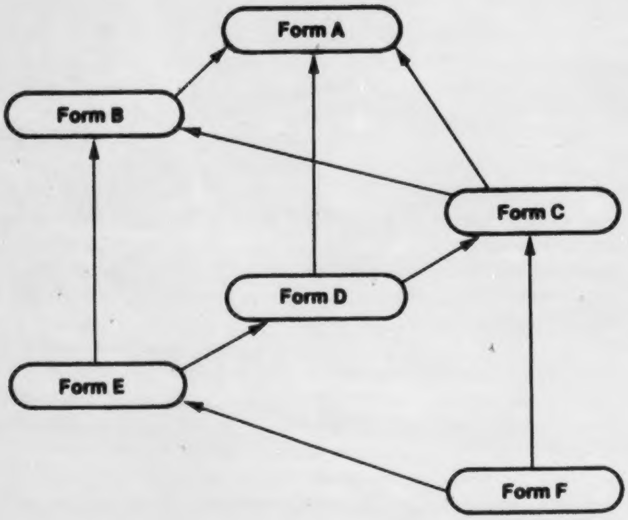
$$A = \frac{A_1 \sqrt{1 + A_2^2} + A_2 \sqrt{1 + A_1^2}}{\sqrt{1 + A_2^2} + \sqrt{1 + A_1^2}}$$

and

$$B = \frac{B_1 \sqrt{1 + A_2^2} + B_2 \sqrt{1 + A_1^2}}{\sqrt{1 + A_2^2} + \sqrt{1 + A_1^2}}.$$

The effect of this procedure is that small equating errors are averaged out over a period of time.

Planning of the sequence of double equating involves "braiding," a technique for establishing form-to-form relationships that reduces the danger of developing distinct scale "strands." The equating sequence shown in the diagram below illustrates this principle. For each arrow linking a new form to a previous form, there must be a set of equating items common to both forms.

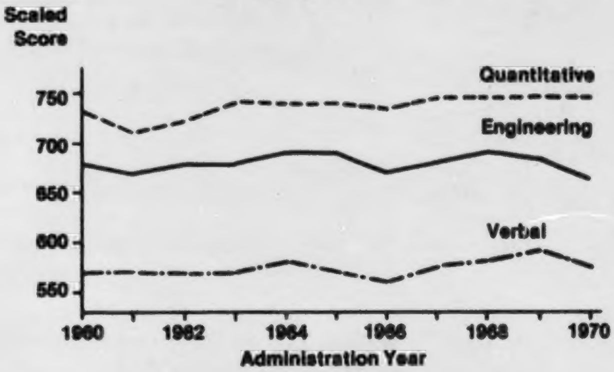


Form A is the first form in the series. Form B is equated to Form A. Then Form C is equated to both A and B, and the bisector of the angle between the two equating lines is used for the final conversion. In the same manner, D is related to both C and A, E is related to B and D, and F is related to E and C. These equating practices are effective in establishing and maintaining a score scale with the following characteristics for each test:

1. The number designating a particular point on the scale represents the same level of development or achievement regardless of the test form on which the score was earned.
2. That number represents the same level of development or achievement for any individual or group of individuals taking the test.
3. That number represents the same level of development or achievement for any form of the test administered over an extended period of time.

The effectiveness of these procedures is illustrated in Figure 1, which shows the mean scaled scores of the National Science Foundation (NSF) Graduate Fellowship applicants in engineering from 1960 to 1970.

Figure 1: Level of Performance of NSF First-Year Engineering Applicants



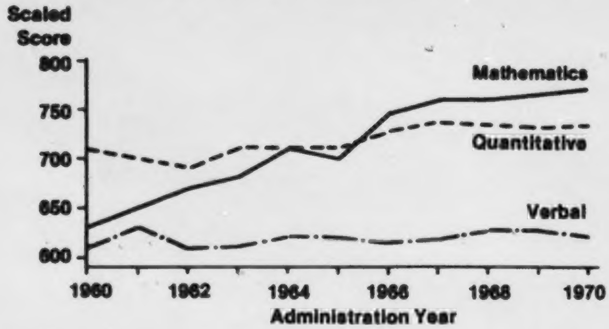
The NSF applicant groups applying for first-year graduate fellowships are similar from year to year. The expectation is that their test performance will exhibit the same characteristics over a 10-year period. The graph shows that their mean scores on the Advanced Engineering Test and the Aptitude Test between 1960 and 1970 have remained as steady as can be expected.

Stability of the Scaled-Score System

There is yet another aspect of scale stability that, although dependent on the stability maintained through equating, can be affected by factors outside the statistical considerations described in the preceding sections. In the description of the development of the GRE scaled-score system, emphasis was placed on the relationship between Aptitude and Advanced Test scores and the fact that the scale for each Advanced Test was designed to reflect the ability level of the population taking the test. Figure 1 shows the stability of that relationship for the Advanced Engineering Test. The engineering applicant group is, as one would expect, a relatively high ability group, and the verbal and quantitative ability score means are well above the GRE average. The mean scores on the Advanced Engineering Test are consistently between the verbal and quantitative means, thus showing the expected relationships among the three scores and demonstrating the stability of the GRE scaled-score system as far as the Advanced Engineering Test is concerned.

The GRE scaled-score system is defined in terms of the basic reference group of 1952 and the educational experience of that group. What happens when the educational experience of more recent groups differs drastically from that of the 1952 group? A few years before 1960, there was a strong movement of mathematics curriculum revision that introduced concepts of modern mathematics into the high school curriculum and then into the elementary school curriculum. The effect of this movement was to speed up the mathematics experience of some students before they reached college and to enable them to do college level work while still in high school. Consistent with that curriculum change is the change in performance on the GRE Advanced Mathematics Test shown in Figure 2, which presents the mean scores of NSF first-year applicants in mathematics.

Figure 2. Level of Performance of NSF First-Year Mathematics Applicants



The relationship between each Advanced Mathematics Test mean and the corresponding Aptitude Test means is near the expected value from 1960 to 1962, but the consistent upward trend in the Advanced Mathematics Test means is not accompanied by a corresponding trend in the Aptitude Test means. From 1964 to 1970 (and continuing to the present), the Advanced Mathematics Test performance is significantly higher than one would expect from the Aptitude Test performance.

Rescaling Study of 1967–68

The trend observed in the case of the NSF first-year mathematics applicants was also evident in the general GRE population for other tests. It was inevitable that long-range factors would change the meaning of the scaled-score system and affect the interpretation of GRE scores. Among the changes that were operating since the establishment of the scale were changes in the GRE population (approximately a threefold increase in Aptitude Test volume from 1964 to 1970), curriculum changes at the high school level (CBAP: Chemical Bond Approach Project; CHEMS: Chemical Education Material Study; BSCS: Biological Sciences Curriculum Study; PSSC: Physical Science Study Committee; SMSG: School Mathematics Study Group; etc.), and changes in the relationships among the GRE verbal ability, quantitative ability, and Advanced Test scores.

By 1968 the accumulated evidence of changes in the meaning of the scaled-score system was sufficient to warrant a statistical investigation. A rescaling study based on the 1967–68 scores was begun in 1968 to determine what had happened to the scale in the 15 years since its development. A report of the results of this study was prepared for the GRE Board (Wallmark, 1969), showing the magnitude of the change for each test. As had been predicted, the major changes occurred in mathematics and the sciences. The results of the study are summarized in Table 16, which shows the means and standard deviations on the Aptitude Test and on the Advanced Test for each Advanced Test group. A comparison of the old-scale and possible new-scale statistics shows the magnitude of the adjustment that would result from a rescaling. Note in particular the decreases in standard deviations for the Advanced Mathematics and Physics Tests.

Table 16: Scaled Score Means and Standard Deviations of the 1967-68 Rescaling Samples

Advanced Test Subgroup	N	Verbal Ability		Quantitative Ability		Advanced Test			
						(Old Scale)		(Possible New Scale)	
		Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Biology	4,696	539	110	580	104	614	107	544	106
Chemistry	2,486	548	119	663	92	613	100	588	102
Economics	1,930	552	118	618	107	621	111	567	110
Education	2,746	472	95	465	104	471	74	476	99
Engineering	4,259	503	119	697	72	619	102	624	91
French	1,292	586	106	512	105	559	88	556	103
Geography	306	511	101	534	107	502	86	516	102
Geology	575	540	110	615	93	581	[illegible]	[illegible]	102
History	4,919	567	109	515	110	554	80	[illegible]	104
Literature	6,276	607	101	516	109	570	85	580	100
Mathematics	3,279	558	119	700	81	653	150	632	97
Music	647	519	113	502	114	518	91	508	107
Philosophy	793	625	104	573	119	649	118	586	104
Physics	2,190	578	119	712	73	614	138	632	94
Political Science	2,745	568	110	536	113	519	90	542	106
Psychology	5,643	562	104	559	109	544	89	546	104
Sociology	2,151	537	119	502	123	533	116	522	116
Spanish	770	545	111	479	104	545	101	518	101
Speech (Total)	695	513	95	466	101	458	85	502	99
Entire Scaling Group	48,398	554	116	574	128				

The rescaling study demonstrated that the across-field comparisons of Advanced Test scores that were possible with the 1952 senior population could not be made with the 1967–68 senior population; and such comparisons cannot be made with a current group. The GRE Board considered the implications of the study of the scale and deliberated over the possibility of rescaling the Advanced Tests. Rescaling would have permitted users to make limited comparisons of Advanced Test scores in different fields. However, rescaling would also have meant that past scores and rescaled scores for a given Advanced Test would not have the same meaning and could not be compared. The issue was fundamentally this: Which will be more beneficial to most users—the ability to compare the Advanced Test scores of people in different fields taking different Advanced Tests or the ability to study score trends effectively within a field? Although fellowship sponsors might find comparisons of Advanced Test scores across fields to be important, most institutions using the scores would be interested in comparing the scores of students taking the same Advanced Test and would find comparability across years to be essential to study of the usefulness of the scores over a period of time. Thus, it was decided to continue use of the original scale, to meet the needs of the majority of users, and to provide rescaling information to fellowship sponsors requiring the capacity for across-field comparisons of Advanced Tests to supplement the across-field comparisons already possible using the Aptitude Test. Thus, the *Guide to the Use of the Graduate Record Examinations* instructs users that scores across Advanced Tests are *not* comparable although scores from one year to the next for a given Advanced Test are comparable. Although the linkage of the Advanced Tests with the Aptitude Test has not been sustained, for historical reasons any newly introduced Advanced Test is scaled by the same method used in the past, providing initial linkage in the year of introduction.

Reliability and Error of Measurement

Reliability is the extent to which a test is consistent in measuring whatever it does measure. It indicates how much of the variation in the results of testing a group of individuals can be attributed to the systematic sources of variation one is trying to measure and how much to other sources of variation that may be classed as errors of measurement. The index of reliability is usually stated as a correlation coefficient for two sets of similar measurements and may be interpreted as the ratio of the true-score variance to the observed-score variance.

One method of estimating the reliability of a test is to administer the test twice to the same group of individuals and to correlate the two sets of measurements. Because this method has both statistical and practical disadvantages, it is seldom used. For the GRE this procedure would introduce memory as an unwanted factor of systematic variation, thus overestimating the reliability. The main practical disadvantage for GRE is that the three-hour time limit of each testing session is excessive for a double session and would introduce a factor of fatigue.

The statistical disadvantage of the test-retest method can be overcome by using an alternate parallel form of the test for the second testing. Although this second method does not solve the problem of the practical disadvantages mentioned above, it is the preferred method when speed is an important factor in the scores.

The split-halves method, a variation of the parallel-forms method, can be used for estimating the reliability of tests that are sufficiently long. One form of the test is administered to a group of individuals. The test is divided into halves for scoring, and the correlation of the two sets of scores is used as the reliability estimate after correction for test length. The difficulty with this method lies in splitting the test into halves that can be considered parallel in both content and statistical properties. The main advantage, however, is that the effects of content sampling are considered without the effects of memory or response variation over time.

The method of estimating reliability coefficients used in the GRE Program employs analysis of variance procedures with a single administration of one test form. This method, proposed by Kuder and Richardson (1937) and further developed by Dressel (1940), is based on interitem correlations and lends itself well to computer processing. The formula, now known as Kuder-Richardson formula (20), is given below:

$$r_{tt} = \frac{n}{n-1} \cdot \frac{\sigma_t^2 - n\overline{pq}}{\sigma_t^2},$$

where

r_{tt} = the estimated reliability of test t,
σ_t^2 = the observed variance of test t,
n = the number of items in test t,

and

$\overline{pq}$ = the average of the item variances.

This Kuder-Richardson formula (20) has been adapted for formula scoring (see page 43). It provides an average of all reliability coefficients that can be obtained from all possible ways of splitting the test.

There has been considerable debate among measurement specialists as to the appropriateness of the Kuder-Richardson formula (20) in estimating the reliability of a test when all examinees do not finish the test. The criticism is important for the GRE Program because some of the Advanced Tests show moderate speededness when the usual criteria are applied. However, a good portion of the "speededness" may reasonably be attributed to difficulty. Frances Swineford (1973) made a comparison of the Kuder-Richardson formula (20) results obtained on moderately speeded forms of the College Board Scholastic Aptitude Test with results obtained by using three other methods and demonstrated that the Kuder-Richardson formula (20) can be used with confidence for estimating the reliability of tests that are moderately speeded.

Reliability coefficients indicate the portion of score variance that can be attributed to true-score variation, but it is of limited value for test users. The standard error of measurement is the statistic commonly used to interpret scores because it is a measure of reliability stated in score units and indicates the probable range of discrepancy between the observed scores and the true scores of a group. In interpreting test scores for a group, one may say that the true score for each individual in the group lies within an interval extending from one standard error above to one standard error below the observed score and expect to be right about 2 times out of 3. By extending the range to two standard errors above and below the observed score, one increases the proportion of correct statements to 95 out of 100.

Item and Test Analysis

Equating is based on the assumption that the various forms of a test are parallel in both content and difficulty. To the extent that this condition is not met, the standard error of measurement computed from a distribution of scores earned on a number of different forms would be higher than that estimated from scores on only one form of the test. It is obvious, therefore, that achieving parallelism is an important aspect of test construction, and quality control procedures must be established to evaluate the success of that effort. The specifications for any given test within the GRE Program consist of three principal elements: 1) the distribution of item content and skills to be assessed; 2) the distribution of item difficulty; and 3) the distribution of item-test correlations. This section is concerned with the two latter elements and with other psychometric properties of the test.

Item Analysis

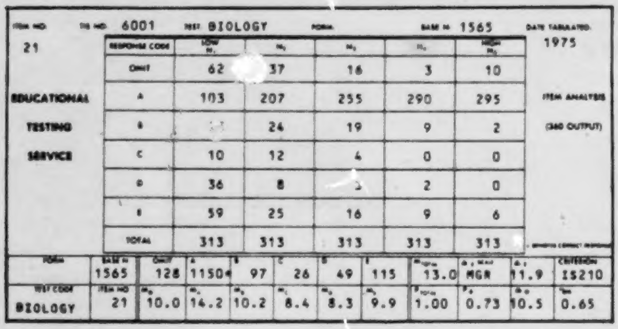
Item analysis is a statistical procedure that provides detailed information about each item, describing the relative attractiveness of the options, the difficulty level of the item for the analysis sample, the power of the item to discriminate among the examinees with respect to a given criterion, and the way the item functions in a particular test.

Before an item is used in a final form of the Aptitude Test, it is pretested and analyzed to identify possible weaknesses and to determine difficulty level and discrimination. Items that pass inspection are then placed in an item pool to make them available for test assembly; those that do not pass are rewritten to correct the flaws or are discarded. In the GRE Program, pretesting serves a useful quality control function for Aptitude Test construction.

For the Advanced Tests, although analyzed items are used as equators, reliance is placed on the subjective judgment of the committee of examiners in estimating difficulty level and discriminating power of newly written items. To a remarkable degree, this procedure is effective. Nevertheless, when a new form is first administered, it is subjected before final scoring to a preliminary item analysis for the sole purpose of identifying items that may be ambiguous or faulty. For most forms no problems emerge. When a problem is identified through the analysis and is confirmed by the subject matter specialists, the item is dropped from scoring if there is no correct response or double-keyed if there are two correct responses. Since GRE Advanced Tests contain an ample quantity of items, reliability remains above .90 in spite of the deletion.

Normalized Total-Group Method of Item Analysis. The method of item analysis used in the GRE Program is called the normalized total-group method. A sample analysis of an Advanced Biology Test item is shown in Figure 3. The item analysis form consists of two parts, the upper grid showing the tally of how many examinees in each quintile group selected each distracter, and the lower strip providing the essential item statistics.

Figure 3: Item Analysis Sample



ITEM NO. 21 TIS NO. 6001 TEST BIOLOGY FORM SAM N 1565 DATE TABULATED 1975

EDUCATIONAL TESTING SERVICE

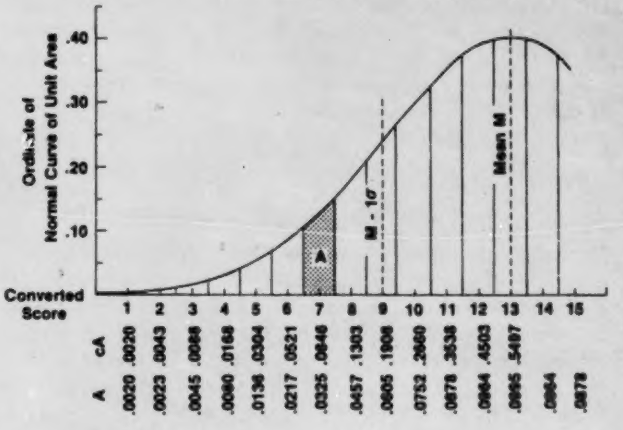
ITEM ANALYSIS (360 OUTPUT)

RESPONSE CODE	LOW				HIGH
OMIT	62	37	16	3	10
A	103	207	255	290	295
B	[illegible]	24	19	9	2
C	10	12	4	0	0
D	36	8	3	2	0
E	59	25	16	9	6
TOTAL	313	313	313	313	313

FORM	SAM N	OMIT	A	B	C	D	E	M TOTAL			CRITERION
	1565	128	1150	97	26	49	115	13.0	MGR	11.9	IS210
TEST CODE	ITEM NO							P TOTAL	P+		
BIOLOGY	21	10.0	14.2	10.2	8.4	8.3	9.9	1.00	0.73	10.5	0.65

Under optimum conditions, the item analysis sample is selected to be representative of the population for which the test was designed and to be adequate in size. The criterion selected for the analysis is usually the total score on the test or subtest, but may be an external criterion related to the item type. The distribution of criterion scores is then converted to a normal distribution with a mean set at 13.0 and a standard deviation at 4.0. The transformation is shown in Figure 4, which shows the left half of a normal curve divided into sections one-fourth standard deviation wide. Each section represents a converted score, and the area of that section indicates the proportion of observations in the interval identified on the graph by the value of its midpoint. For example, if the sample size is 1,000, the two lowest scores will be converted to 1 (1,000 × .0020), the next two will be converted to 2 (1,000 × .0023, rounded), and so on. The cumulative area cA is used to resolve rounding problems. Thus, each individual in the sample is assigned a converted score in the range of 1 to 25, and this score is used in the item analysis computation.

Figure 4: Criterion Score Conversion



The item analysis sample is divided into five equal subgroups based on the criterion score: the lowest fifth, the second fifth, the middle fifth, and so on. The responses of the sample are tallied and the frequencies entered in the grid of the item analysis form. This display is used to determine how each option functions in discriminating among the examinees. The total frequency for each response and for omissions is entered in the strip at the bottom of the form with the mean converted criterion score for that group. In Figure 3, for example, 128 examinees omitted the item, and this group has a mean criterion score of 10.0; 1,150 chose the correct response A and have a mean criterion score of 14.2. The right-hand portion of the form shows the mean criterion score of the total group (M_{total}) to be 13.0, the proportion of the group reaching the item (P_{total}) to be 1.00, and the proportion answering correctly (P_+) to be .73 P_+ is translated into a difficulty index, Δ_o, which is also on the 13.0/4.0 scale. The biserial correlation is determined by the formula

$$r_{bis} = (M_+ - M_{total}) \times \frac{1}{\sigma}\frac{P_+}{y}$$

where

M_+ is the mean criterion score for the group answering correctly,

σ is the standard deviation for the total sample (set equal to 4.0 in this method),

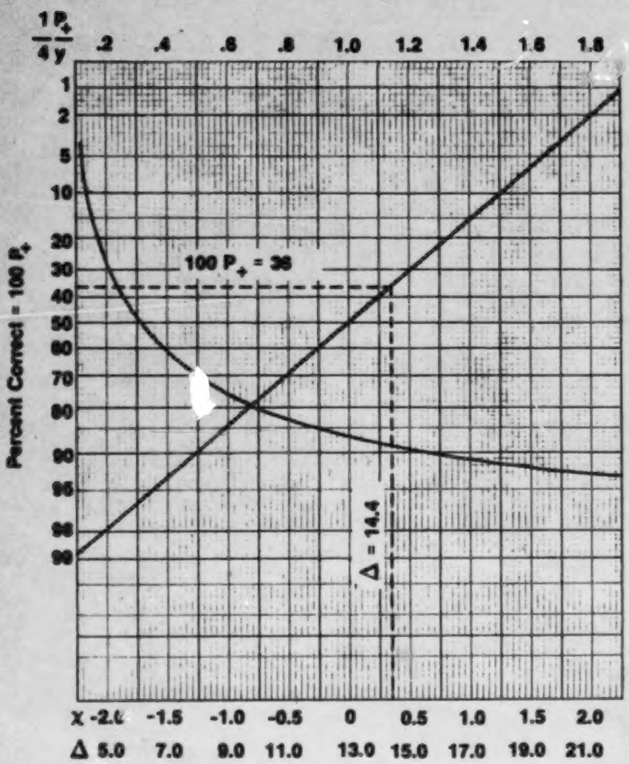
P_+ is the proportion answering correctly,

and

y is the ordinate in the unit normal distribution which divides the area under the normal curve into P_+ and $1 - P_+$.

The relationships between P_+ and Δ and between P_+ and $\frac{1}{4}\frac{P_+}{y}$ when M_{total} is 13.0 are shown in Figure 5.

Figure 5: Relationship between P+ and Δ when $M_{total} = 13.0$



For tests that tend to be difficult for the group, not all examinees find time to attempt all items. An assumption is made that an individual has attempted every item up to and including the last one that he or she answered and did not reach any of the remaining items. In the typical case, the dropping out begins in the lowest ability group, with a consequent change in the ability level of the group that does reach the item. M_{total} is based on the scores of those who, by this definition, reach the item. This is not a perfect description of what occurs, but, provided that the proportion of examinees who drop out is not extremely large, it is a reasonable basis for using a variable M_{total} in computing Δ_o when speed and power are highly correlated, as is the case in the GRE tests.

The dropping out is indicated in two ways on the item analysis form; the last row of the grid shows the number of examinees in each group who have reached the item, and the P_{total} box in the strip shows the proportion of the total sample reaching the item. M_{total} is the mean criterion score for that group and usually increases toward the end of the test. The difficulty index is then computed by the formula

$$\Delta_o = M_{total} + 4x,$$

where

x = the deviation from the mean of the normal curve corresponding to P_+, stated in standard deviation units, as shown in Figure 5.

The factors that affect r_{bis} include: the degree of independence of the item and the criterion (item included in the score or item not included), the nature of the criterion (same subject matter as the item and homogeneous, different subject matter, or mixed subject matter), and the range of the raw criterion (less than 25 raw-score points or greater). To aid in interpreting the biserial correlation for a particular analysis, a coded description of the criterion score is included in the upper right-hand box of the item analysis strip. There are three parts to the code: the first letter indicates the location of the item with respect to the criterion (I means internal, X means external); the second letter indicates the nature of the criterion (S means that the subject matter is homogeneous and the same as that for the item, D means that the subject matter is different, and M means that it is mixed); and the number indicates the number of items in the criterion (If the item is included and the criterion is based on fewer than 25 items, the biserial will have a spurious component).

In the GRE Program, a recurring problem with item analysis is that the sample available for item and test analysis is not representative of the total group for which the test was designed because groups taking the test vary by year and, within a year, by administration date. If the sample is very able, the observed deltas will be relatively low. If two forms of a test have been analyzed with samples of different ability, the observed deltas cannot be directly compared. This problem is solved by establishing a base scale with a sample selected to be representative of applicants for admission to graduate school. In the GRE Program, the basic reference sample for item analysis for each test consists of the seniors who were tested in the academic year 1962–63. The samples for the analyses of the next new forms, which were introduced in 1964–65, were selected to represent the reference groups and were used to establish the Δ scales. In all subsequent analyses, the observed deltas, Δ_o, were equated to place them on the base scale, and the equated values, Δ_E, were entered on the item analysis form along with the scale identification.

Delta equating is accomplished by relating the observed results of an item analysis of a test to the standard reference population by including in the test a group of items that have been used in the program and for which equated deltas are already known. This set of items can be, and usually is, the set used for score equating where common items are used for equating. A scatterplot relating the observed deltas obtained in the new analysis with the corresponding equated deltas from a previous analysis is then used to generate an equating line. This line is then used to estimate equated deltas for the new items in the test.

$$\Delta_E = a\,\Delta_o + b,$$

where $a = \sigma_{\Delta_E}/\sigma_{\Delta_o}$, and $b = M_{\Delta_E} - aM_{\Delta_o}$.

Test Analysis

Test analysis is the final stage in the test development process. When a final form of a test has been administered, it is subjected to a detailed analysis to determine the extent to which the test specifications have been met and to provide information that can be used to guide future test construction. The analysis data are used to evaluate the test as a whole, to determine its efficiency, its discrimination in the score region used for selection decisions, the reliability of the reported scores, the intercorrelations among reported scores and special subscores, its speededness characteristics, the distributions of item difficulty indexes and biserial correlations for the reported scores and special subscores, and special score data to study examinees' patterns of response to test questions. All this information is condensed in a test analysis report, written as a within-office quality-control document providing information for test specialists and test committees, for the research staff, and for clients and test experts who review the tests.

The first step in the planning of a test analysis is to select an appropriate sample. Under ideal conditions, the analysis sample would represent the population for which the test was designed. Under practical conditions this ideal is seldom possible. Although the sample may not represent the target population, it almost always represents the total group examined at the first administration of the test. The sample size should be sufficiently large to ensure reliable results. Between 800 and 1,000 cases in the sample is considered adequate, but, if the total group tested at the first administration is less than 3,000, all cases are generally used in the sample. If the total group is much larger than that, a sample is selected by taking one case out of every n cases to ensure a sample size of approximately 1,000 cases. If the total group is much smaller than 800, a decision is made whether to do the analysis with the available cases or to delay until additional cases from some future administration of that form become available and can be combined with the original group.

The data processing phase of the test analysis work produces the following computer output: distributions for all reported scores and item analysis criterion scores, item analysis, a delta vs. r-biserial distribution sheet for each set of items analyzed, the results of the delta equating, an intercorrelation table of all variables used in the analysis, and a test analysis tabulation for each separately timed section of the test. This detailed information is then analyzed and summarized in the test analysis report.

Examples of the test analysis data for a recent form of the GRE Advanced French Test are presented in Tables 17 to 24. Table 17 presents two frequency distributions, one based on the performance of the total group of examinees who took this particular test form when it was first administered in October 1976 and the other based on the performance of the total group examined during the year from October 1975 through September 1976. As can be seen from the first distribution, the number of cases available for the analysis sample is well below the number considered to be adequate. Although that number is small, a comparison of the total-group mean and standard deviation with the corresponding statistics of the 1975–76 norms group shows the total group for October 1976 represents the norms group rather well. This fact, combined with the need for some information to help in the construction of the next form, led to a decision to proceed with the analysis. At the bottom of Table 17 are summary statistics and conversion data, including the conversion equation for translating raw scores to scaled scores.

Table 17: Total Score Distributions

Advanced French Test, Form A

(Taken by candidates for admission to graduate schools, October 1976)

October 1976 Administration				October 1975—September 1976 Norms		
Raw Score X	Standard Score Y	f	Percentile Rank of Lower Limit of Interval	Standard Score	f	Percentile Rank of Lower Limit of Interval
180	770	1	99.5	800 - 820	1	99.9
171 - 179	740 - 770	1	99.0	770 - 790	4	99.6
162 - 170	720 - 740	–	99.0	740 - 760	9	98.8
153 - 161	700 - 720	3	97.4	710 - 730	21	97.0
144 - 152	670 - 690	9	92.8	680 - 700	29	94.5
135 - 143	650 - 670	4	90.7	650 - 670	45	90.6
126 - 134	620 - 650	5	88.1	620 - 640	68	84.8
117 - 125	600 - 620	12	82.0	590 - 610	66	79.1
108 - 116	580 - 600	16	73.7	560 - 580	135	67.4
99 - 107	550 - 570	21	62.9	530 - 550	141	55.3
90 - 98	530 - 550	19	53.1	500 - 520	156	41.9
81 - 89	500 - 530	12	46.9	470 - 490	166	27.6
72 - 80	480 - 500	28	32.5	440 - 460	126	16.7
63 - 71	460 - 480	19	22.7	410 - 430	85	9.4
54 - 62	430 - 450	15	14.9	380 - 400	48	5.3
45 - 53	410 - 430	10	9.8	350 - 370	35	2.2
36 - 44	390 - 410	8	5.7	320 - 340	23	0.3
27 - 35	360 - 380	5	3.1	290 - 310	2	0.1
18 - 26	340 - 360	3	1.5	260 - 280	1	0.0
9 - 17	310 - 330	1	1.0			
0 - 8	290 - 310	2	0.0			
		194			1161	
$M_x = 86.86$ $\sigma_x = 33.25$ $M_y = 521$ $\sigma_y = 89$ $Md_x = 84.50$ (195 items)	Conversion Data Converted to the GRE scale through scores on 40 items in common with one form and 40 items in common with another form. $Y = 2.6612\,X + 289.3151$			$M_y = 519$ $\sigma_y = 90$		

The information in Table 17 is used to describe the efficiency and skewness of the test. Under normal circumstances, a test is most efficient for the group if the score distribution covers the entire possible score range. In this case the maximum obtained formula score of 180 is nearly one-half standard deviation below the maximum possible score of 195. This characteristic is common among the GRE Advanced Tests, which cover subject matter selected from a broad range of undergraduate curricula rather than from one universal curriculum. The test was relatively difficult for the group. The mean score on the 195 questions is 86.86. A test of middle difficulty would be expected to yield a mean formula score equal to approximately half the total number of test questions, in this case 92.5, the score that would be expected for an examinee who knew the answers to half the items and responded at random to the remaining ones. The skewness of the score distribution is another characteristic used for evaluating the effectiveness of the test construction. Skewness, or the third moment, is defined by the formula,

$$\alpha_x = \frac{\sum x_i^3}{N\sigma^3},$$

where x_i is the deviate score $(X_i - M_x)$ of the i^{th} examinee and the summation is done over the number of examinees (N). A convenient approximation of skewness is given by the formula $3(M - Md)/\sigma$, which in this case is .213. This estimate is not reliable for a sample size as small as the N for the total group, but, it does indicate some positive skewness, which usually means that the test is relatively difficult for the group tested.

Two subscores are reported for the Advanced French Test. The distributions of these scores for the total group are presented in Table 18. Using the same analysis procedures, one can see that the Interpretive Reading Skills subtest is rather easy for the group and that the score distribution is characterized by a high mean and slightly negative skewness. The Literature and Civilization subscore, on the other hand, has a very low mean and is positively skewed, indicating a very difficult subtest. With this information,

Table 18: Subscore Distributions

Advanced French Test, Form A
(Taken by candidates for admission to graduate schools, October 1976)

1: Interpretive Reading Skills				2: Literature and Civilization			
Raw Score X	Standard Score Y	f	Percentile Rank of Lower Limit of Interval	Raw Score X	Standard Score Y	f	Percentile Rank of Lower Limit of Interval
90 - 94	72 - 74	1	99.5	90 - 94	78 - 80	1	99.5
85 - 89	69 - 71	4	97.4	85 - 89	76 - 78	1	99.0
80 - 84	67 - 69	3	95.9	80 - 84	73 - 75	–	99.0
75 - 79	65 - 67	8	91.8	75 - 79	71 - 73	2	97.4
70 - 74	62 - 64	9	87.1	70 - 74	68 - 70	7	94.3
65 - 69	60 - 62	12	80.9	65 - 69	66 - 68	9	89.7
60 - 64	57 - 59	22	69.6	60 - 64	63 - 65	6	86.6
55 - 59	55 - 57	24	57.2	55 - 59	61 - 63	13	79.9
50 - 54	52 - 54	16	49.0	50 - 54	58 - 60	9	75.3
45 - 49	50 - 52	21	38.1	45 - 49	56 - 58	12	69.1
40 - 44	47 - 49	24	25.8	40 - 44	53 - 55	17	60.3
35 - 39	45 - 47	16	17.5	35 - 39	51 - 53	24	47.9
30 - 34	42 - 44	10	12.4	30 - 34	48 - 50	15	40.2
25 - 29	40 - 42	5	9.8	25 - 29	46 - 48	22	28.9
20 - 24	37 - 39	6	6.7	20 - 24	43 - 45	24	16.5
15 - 19	35 - 37	6	3.6	15 - 19	41 - 43	22	5.2
10 - 14	32 - 34	1	3.1	10 - 14	39 - 40	5	2.6
5 - 9	30 - 32	4	1.0	5 - 9	36 - 38	4	0.5
0 - 4	28 - 30	2	0.0	0 - 4	34 - 36	–	0.5
				–5 - –1	34	1	0.0
		194				194	
$M_x = 49.64$ $\sigma_x = 17.95$ $M_y = 52.0$ $\sigma_y = 8.9$ $Md_x = 50.50$	Conversion Data Converted to the GRE scale by setting the mean and standard deviation equal to one-tenth of the total-score mean and standard deviation, respectively.			$M_x = 37.31$ $\sigma_x = 17.84$ $M_y = 51.9$ $\sigma_y = 8.9$ $Md_x = 35.00$	Conversion Data Converted to the GRE scale by setting the mean and standard deviation equal to one-tenth of the total-score mean and standard deviation, respectively.		
(92 items)	$Y = 0.4930\,X + 27.5637$			(103 items)	$Y = 0.4962\,X + 33.5380$		

Table 19A: Scoring Formulas and Reliability Coefficients

Advanced French Test, Form A
(Description of Sample: Total group to highest multiple of 5; N = 190.)

Score	Number of Items	Scoring Formula	Reliability*	SE_{meas} Raw Score	SE_{meas} Scaled Score
1. Interpretive Reading Skills	92	R-W/4	.926	4.89	2.41
2. Literature and Civilization	103	R-W/4	.916	5.09	2.53
3. Total Score	195	R-W/4	.955**	7.06	18.78

*Adaptation of Kuder-Richardson formula (20).
**See text.

Table 19B: Intercorrelations

Score	1	2	3	Mean	S.D.
1. Interpretive Reading Skills	—	.742	.935	49.61	17.97
2. Literature and Civilization	.742	—	.932	36.94	17.58
3. Total Score	.935	.932	—	86.45	33.11

Table 19C: Speededness of Test

Percent completing test	61.6
Percent completing 75 percent of test	100.0
Number of items reached by 80 percent of the candidates	194
Total number of items	195

the committee of examiners determines whether or not the high difficulty level is consistent with their knowledge of French literature curriculum design and emphasis.

Table 19 presents three kinds of information: the estimated reliability of the reported scores, the correlations among the reported scores, and the speededness of the total test. The reliability of each subscore was estimated by the Kuder-Richardson formula (20) adapted by Dressel (1940) for use with formula scoring:

$$\text{rel.} = \frac{n}{n-1}\left[1 - \frac{\sum pq + k^2 \sum p'q' + 2k \sum pp'}{\sigma_t^2}\right],$$

where

n = the number of items,
p = the proportion of examinees answering correctly,
$q = 1 - p$,
p', q' = the corresponding proportions for incorrect responses,
k = the correction factor in the formula scoring,
σ_t = the standard deviation of the score distribution,

and the summations are over n items. The corresponding formula for the standard error of measurement, which translates the reliability information into a statistic measured in score units, is:

$$SE_{meas} = \sigma_x \sqrt{1 - \text{reliability}}$$

The estimated reliability of the total score is computed by the formula,

$$\text{reliability} = 1 - \frac{\text{error variance}}{\text{total variance}},$$

where the error variance is the sum of the squared standard errors of measurement of the two subscores, and the total variance is the variance of the total score on the test. The total-score reliability obtained by this formula is normally higher than that obtained by applying the K-R (20) formula to the total test, but the difference is small when the test is homogeneous. In the case of the Advanced French Test (Table 19), the two estimates are .9546 and .9540, respectively. If a score is a factor in a decision that affects the future educational goals and career choices of an individual, the estimated reliability of that score should be high (and the corresponding standard error of measurement low). The standard commonly used in the GRE Program is that the reliability should be at least .90 for the total score on any test. The subscores, which are not intended for selection purposes, may have somewhat lower reliability estimates, with lower limit set at about .80. For the example in Table 19, all the reliability estimates exceed .90. The standard errors of measurement are given in both raw-score and scaled-score units. To translate the raw-score standard error of measurement into scaled-score units, multiply by the conversion parameter A, which is the slope of the conversion line. The use of the standard error of measurement is explained in the *Guide to the Use of the Graduate Record Examinations*, which is supplied to all institutions using the GRE score reporting service.

Whenever a test generates more than one score for reporting, or has separately timed sections, an intercorrelation table is presented in the test analysis report. Table 19 shows such a table for the Advanced French Test. The correlation of each subscore with the total score is spuriously high because the subscore is included in the total score and is a large part of it. The correlation between the two

Table 20: Score Distributions

Advanced French Test, Form A

TEST Graduate Record Examinations — SUBJECT Advanced French — FORM 1976
SECTION Total Test — ITEMS 195 (5-choice) — DATE

R + W / SCORE	44 TO 52	53 TO 61	62 TO 70	71 TO 79	80 TO 88	89 TO 97	98 TO 106	107 TO 115	116 TO 124	125 TO 133	134 TO 142	143 TO 151	152 TO 160	161 TO 169	170 TO 178	179 TO 187	188 TO 195	SCORE f	R (RIGHT) Number of Items	R (RIGHT) f	W (WRONG) Number of Items	W (WRONG) f	O (OMIT) Number of Items	O (OMIT) f	N R (NOT REACHED) Number of Items	N R (NOT REACHED) f
																			175-184	2			144-152	1		
172-182																	2	2	165-174	-	128-135	2	135-143	-		
161-171																		-	155-164	6	120-127	3	126-134	-	42- 44	1
150-160														1		3	1	5	145-154	9	112-119	3	117-125	1	39- 41	-
139-149														1	2	2	3	8	135-144	9	104-111	4	108-116	1	36- 38	-
128-138															1	2	3	6	125-134	14	96-103	7	99-107	2	33- 35	-
117-127													1	1	1	6	3	12	115-124	20	88- 95	5	90- 98	2	30- 32	1
106-116													3	1	5	5	4	18	105-114	25	80- 87	9	81- 89	7	27- 29	-
95-105											1	2	3	9	5	2	2	24	95-104	24	72- 79	16	72- 80	6	24- 26	-
84- 94									1	1	1	4	5	3	3	1	1	20	85- 94	29	64- 71	16	63- 71	7	21- 23	2
73- 83							2	1	1	4	1	7	4	2	2	4	3	31	75- 84	19	56- 63	31	54- 62	10	18- 20	-
62- 72						1	1	2	2		3	2	3	4	3		2	23	65- 74	13	48- 55	33	45- 53	15	15- 17	-
51- 61						1			2	1	1	4	2		1	2	3	17	55- 64	9	40- 47	23	36- 44	23	12- 14	2
40- 50				1				1	1		1			1	1		1	7	45- 54	7	32- 39	14	27- 35	27	9- 11	1
29- 39			1			1	1	2		1				1	1	2	1	11	35- 44	2	24- 31	13	18- 26	27	6- 8	4
18- 28												1				1	1	3	25- 34	1	16- 23	7	9- 17	25	3- 5	8
7- 17	1						1					1						3	15- 24	1	8- 15	4	0- 8	36	0- 2	171
TOTAL	1	-	1	1	-	3	5	6	7	7	8	21	21	24	25	30	30	190	190		190		190		190	
SCORE R-.2500W								ITEMS									MEAN	86.45	101.07		58.95		33.52		1.46	
																	STANDARD DEVIATION	33.11	29.82		24.47		27.04		4.75	

subscores is not affected in this way because the subscores are independent of each other. If two subtests are parallel, the correlation between the scores will approach the geometric mean of their reliability estimates. Expressed in terms of correction for attenuation, $r_{ab}/\sqrt{r_{aa}\, r_{bb}}$ approaches 1. In this expression, r_{ab} is the correlation between the two subscores and r_{aa} and r_{bb} are the respective reliability estimates. For the example in Table 19, this value is about .81. For a value less than .90 to .95, one may conclude that the two subtests are tapping somewhat different traits or abilities.

At the bottom of Table 19 are data relating to speededness. The guideline at ETS is to regard a test as essentially unspeeded if at least 80 percent of the group reach the last item and if virtually every one reaches at least three-quarters of the items. Any conclusion based on this information must be supported by the special score data presented in Table 20.

Table 20 shows five distributions: the total formula-score distribution and distributions of the number of items answered correctly, answered wrong, omitted, and not reached. It also shows a scatterplot, or two-way table, of the Score versus R + W. Two lines appear on the scatterplot, a solid line that passes through equal values of Score and R + W, and a dashed line near the bottom of the plot that marks off the "chance" area. If speed is an important element in the score, the "not reached" mean and standard deviation are high relative to the corresponding statistics of the score distribution, and there is a noticeable tendency for the entries in the scatterplot to lie close to the diagonal solid line. If power is important, the "not reached" statistics are relatively low, and the entries in the scatterplot cluster in the right-hand columns. In the example shown in Table 20, the evidence points to a measure of power rather than speed.

The dashed line near the bottom of the scatterplot is used to describe the efficiency of the test. If an answer sheet were marked at random, the resulting score would be expected to approximate zero, and the chances are 99 out of 100 that a score obtained in this manner would lie below the dashed line. An efficient test would not produce a large proportion of scores in this chance area. For the example shown in Table 20, this proportion is less than .01.

Table 21 presents two sets of distributions, one set based on the observed deltas of the reported scores and the other on the biserial correlations of items with the total score. For a test of middle difficulty for the group, the mean delta would be about 12.0. In the example, the mean delta for the total test is slightly above this value. The difficult questions seem to concentrate in the second subscore. The deltas have been equated to put them on the reference scale established in 1963 and defined by the reference sample of seniors selected at that time. Delta equating is done so that the equated deltas obtained with a current sample can be compared directly with equated deltas from previous forms of the test. The adjustment resulting from this equating is in this case very slight.

Normally, the criterion used in the analysis is the score on the set of items analyzed. The criterion for the total test would be the total score, and that for each subtest would be the subscore itself. In this case, the test specialist requested that the total score be used throughout. If the subscores had been used as criteria, the mean r_{bis} for each subtest analysis would have been higher. The recommendation for the committee of examiners would be that it review carefully the seven items with biserial correlations below .20. Such low values are unusual in a language test and may indicate faulty items or items based on subject matter most students do not encounter in their undergraduate programs. Although the item distribution sheet shown as Table 22 is not normally included in the test analysis report, it is used by the committee of examiners to judge the appropriateness of an item that may have questionable statistics. If, for example, an item has a low r-biserial and a low delta, it produces a tally in the lower left-hand portion of the item distribution sheet, thus indicating that the item is very easy for the group of examinees sampled and does not discriminate well among them. On the other hand, if the r-biserial is low and the delta is very high, this is sufficient evidence of a problem that warrants further investigation.

In the course of examining the data on the most recent form of a test to prepare for the construction of a new form, the committee of examiners usually compares the analysis results of the five most recent forms. A summary of the results for the total scores appears as Table 23. A similar table for the subscores appears as Table 24. Thus, the test analysis report is a compact but complete record of the essential statistical characteristics of the test and serves as a guide for future test construction.

Descriptive Statistics

The primary reason for providing descriptive data is to help score recipients interpret scores. The usefulness of a test score is

Table 21: Frequency Distributions of Original Deltas and Biserial Correlations, by Score

Advanced French Test, Form A

Delta	Total Score		Interpretive Reading Skills		Literature and Civilization	
18.0-18.9	2				2	
17.0-17.9	2				2	
16.0-16.9	14		2		12	
15.0-15.9	20		4		16	
14.0-14.9	22		6		16	
13.0-13.9	35		12		23	
12.0-12.9	34		24		10	
11.0-11.9	27		16		11	
10.0-10.9	13		10		3	
9.0- 9.9	12		6		6	
8.0- 8.9	6		6			
7.0- 7.9	5		4		1	
6.9 down	3		2		1	
Total n	195	Equated	92	Equated	103	Equated
Mean	12.7	12.6	11.7	11.7	13.7	13.6
σ	2.4	2.2	2.2	2.0	2.2	2.0
a	0.92					
b	0.92					
r_{bis}	Criterion = Total Score					
.80-.89	1		1			
.70-.79	4		4			
.60-.69	17		12		5	
.50-.59	44		21		23	
.40-.49	57		28		29	
.30-.39	44		14		30	
.20-.29	20		9		11	
.10-.19	5		2		3	
.00-.09	1				1	
Negative	1				1	
Not Computed	1		1			
Mean	.44		.47		.40	
σ	.13		.14		.12	

Table 22: Item Distribution Sheet

Advanced French Test, Form A

Raw Delta

r_{bis}	5.9 DOWN	6.0 to 6.9	7.0 to 7.9	8.0 to 8.9	9.0 to 9.9	10.0 to 10.9	11.0 to 11.9	12.0 to 12.9	13.0 to 13.9	14.0 to 14.9	15.0 to 15.9	16.0 to 16.9	17.0 to 17.9	18.0 to 18.9	19.0 UP	Total
*	1															[1]
.80-.89				1												1
.70-.79			2					1		1						4
.60-.69			1		2	3	4	4	3							17
.50-.59		1		2	3	2	5	8	11	2	6	4				44
.40-.49			2		5	4	8	13	7	7	4	7				57
.30-.39				1		2	6	6	8	9	7	2	1	2		44
.20-.29		1		1	1	2	3	2	6	3	1					20
.10-.19				1	1		1				1		1			5
.00-.09											1					1
Neg.												1				1
Total	1	2	5	6	12	13	27	34	35	22	20	14	2	2	0	195

	N	SUM	SUM OF SQUARES	MEAN	S.D.
r_{bis}	194	84.48	40.2441	0.4555	0.1335
Δ_0	195	2481.8	32727.56	12.7271	2.4196

*When P+ is greater than .95, the biserial correlation is not computed.

Table 23: Summary Statistics for Total Score

Advanced French Test

Form Administration Test Analysis Sample N	V October 1972 420	W December 1972 450	X April 1973 360	Y October 1974 370	A October 1976 190
Raw-Score Information					
Number of Items	190	190	190	190	195
Maximum Obtained	159	154	161	150	180
Minimum Obtained	0	3	– 2	66.03	7
Mean	81.60	78.35	69.47	– 3	86.45
S.D.	28.48	24.68	29.12	30.34	33.11
Median	81.67	79.75	67.83	63.63	83.50
Scaled-Score Information					
Mean	540	534	523	517	519
S.D.	83	78	91	91	88
Maximum Possible	860	880	900	890	810
Maximum Obtained	770	770	810	770	770
Minimum Obtained	300	290	310	320	310
Minimum Possible*	300	280	310	320	290
No. of 990 Scores					
Current Norms					
Mean	542	542	542	533	533
S.D.	92	92	92	88	88
Item Statistics					
Mean P+	51.8	51.0	45.7	42.5	52.7
Mean Δ_0	12.8	12.9	13.4	13.8	12.7
S.D. Δ_0	2.7	2.9	2.5	2.4	2.4
Mean Δ_E	13.5	13.3	13.5	14.0	12.7
S.D. Δ_E	2.7	2.4	2.3	2.5	2.2
Mean r_{bis}	.38	.35	.39	.39	.44
S.D. r_{bis}	.13	.14	.14	.12	.13
$r_{bis} < .18$	10 items	17 items	9 items	9 items	5 items
Test Statistics					
Reliability	.940	.921	.942	.947	.955
SE_{meas}					
Raw Score	7.00	6.94	7.01	7.00	7.06
Scaled Score	20.41	21.90	22.00	20.91	18.79
Special Score Data**					
Mean Rights	97.26	95.06	86.28	82.56	101.07
Mean Wrongs	63.24	63.35	67.83	66.67	58.95
Mean Omits	26.17	27.80	29.96	37.74	33.52
Mean Not Reached	3.33	3.79	5.92	3.02	1.46
Speededness**					
% completing test	79 94 85 45	86 41 78 69	54 81 37 51	63.8	61.6
% completing 75%	99 100 99 96	99 100 100 94	99 99 98 90	98.4	100.0
Item reached by 80%	54 55 40 40	55 55 39 37	52 55 40 35	189	194

*The scaled score equivalent to 0 is arbitrarily assigned to negative raw scores.

**Forms introduced in October and December of 1972 and in April of 1973 consist of four separately timed sections. The special score data are based on combined information; the speededness data are given by section.

enhanced when accompanied by relevant information that includes a description of the test and normative and descriptive data that permit evaluation of the performance of an examinee or group of examinees relative to that of an appropriate norms group. The descriptive data for each GRE test are provided in a descriptive booklet made available to students before they take the test and to graduate institutions that use GRE score reporting services. The statistical information, which includes reliability estimates of reported scores, standard errors of measurement, and intercorrelations among reported scores, appears in the *Guide to the Use of the Graduate Record Examinations.* Other kinds of interpretive data based on the performance of groups of students are provided in part in the *Guide*, supplemented by the descriptive statistics in the following sections.

Basic Normative Data

General, or aggregate, norms that provide a broad basis of comparison for graduate institutions consist of percentile rank tables

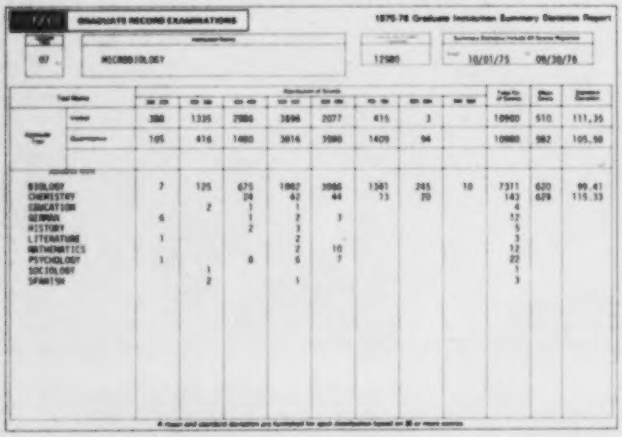
Table 24: Summary Statistics for Subscores

Advanced French Test

Form	V		W		X		Y		A	
Administration	October 1972		December 1972		April 1973		October 1974		October 1976	
Test Analysis Sample N	420		450		360		370		190	
Subscore	Interpretive Reading Skills	Literature and Civilization	Interpretive Reading Skills	Literature and Civilization	Interpretive Reading Skills	Literature and Civilization	Interpretive Reading Skills	Literature and Civilization	Interpretive Reading Skills	Literature and Civilization
Raw-Score Information										
Number of Items	90	100	97	93	92	98	95	95	92	103
Maximum Obtained	83	80	86	68	80	87	80	76	90	92
Minimum Obtained	0	0	4	− 1	− 4	− 2	− 1	− 4	3	− 1
Mean	48.63	33.10	51.61	27.84	42.06	27.53	38.47	27.66	49.61	36.94
S.D.	15.44	15.45	14.85	12.23	16.07	15.01	17.14	15.60	17.97	17.58
Median	50.25	31.38	53.30	26.88	43.05	25.50	38.75	25.36	50.50	34.88
Skewness	high neg.	high pos.	high neg.	high pos.	high neg.	high pos.	(−)	high pos.	mod. neg.	high pos.
Scaled-Score Information										
Mean	54.6	53.0	53.4	53.4	52.3	52.3	51.6	51.8	52.0	51.9
S.D.	8.1	9.0	7.8	7.8	9.1	9.1	8.8	9.3	8.9	8.7
Maximum Possible	76	92	77	95	81	95	81	92	73	85
Maximum Obtained	73	80	71	79	74	88	73	81	72	79
Minimum Obtained	29	34	28	36	28	36	32	35	29	34
Minimum Possible	29	34	26	36	28	36	32	35	28	34
Item Statistics										
Mean P+	63.9	40.6	63.1	38.0	55.6	36.5	49.9	35.2	62.9	43.4
Mean Δ_o	11.6	13.9	11.7	14.2	12.4	14.4	13.0	14.5	11.7	13.7
S.D. Δ_o	2.4	2.3	2.7	2.4	2.3	2.3	2.3	2.2	2.2	2.2
Mean Δ_E	12.3	14.6	12.2	14.3	12.6	14.4	13.2	14.8	11.7	13.5
S.D. Δ_E	2.4	2.3	2.3	2.0	2.1	2.1	2.4	2.3	2.0	2.0
Mean r_{bis}	.44	.40	.42	.35	.43	.38	.42	.42	.47	.40
S.D. r_{bis}	.13	.12	.16	.11	.15	.12	.13	.13	.14	.12
$r_{bis} < .18$ (very low)	4 items	3 items	6 items	7 items	2 items	3 items	3 items	4 items	2 items	4 items
Criterion Score	IRS	LAC	IRS	LAC	IRS	LAC	IRS	LAC	TOTAL*	TOTAL*
Subscore Statistics										
Reliability	.901	.893	.889	.842	.902	.894	.912	.905	.926	.916
SE_{meas}: raw score	4.85	5.05	4.96	4.86	5.03	4.88	5.09	4.81	4.89	5.09
scaled score	2.53	2.94	2.61	3.08	2.86	2.96	2.63	2.87	2.41	2.53

*The test development examiner requested this criterion. If the appropriate subscores had been used as criteria, the mean r_{bis} would have been higher.

Table 25: Frequency Distributions for All 1975–76 Examinees Who Intended to Major in Microbiology



based on the performance of all examinees within a recent three-year norms period. From 1967 to 1977 this type of normative information was the only type provided in the GRE *Guide*, and the percentile ranks that appear with the scaled scores on the score reports are taken from these tables.

The three-year norms have limited value for most graduate institutions. For most users of GRE test scores, the need for identifying an applicant's standing relative to an appropriate reference group is better satisfied by developing local norms based on the institution's own data. In an effort to satisfy this need and to encourage institutions to accumulate local norms, the GRE Program now supplies more detailed information for score interpretation in the form of summary statistics reports based on score data of the most recent reporting year. At the end of a reporting year, each institution that received score reports during that year receives a summary statistics report showing frequency distributions based on all scores reported to the institution in that period, with a count of the number of scores for each test and with the mean and standard deviation for every distribution based on at least 25 cases. Each de-

Table 26: 1970-71 Examinee Volume for the Advanced Tests, by Educational Level

Advanced Test	Percent of Total N						Number of Examinees
	No Response	Jr. Year	Sr. Year	Bachelor's Degree No Grad. Work	1st Yr. Grad. Level	2nd Yr. Grad. Level	
Anthropology	1	4	63	18	6	8	1,190
Biology	1	3	62	20	8	7	13,496
Chemistry	2	3	65	17	7	7	5,126
Economics	2	3	61	18	9	6	4,770
Education	3	1	24	30	28	14	24,179
Engineering	5	3	52	23	11	6	7,858
French	1	3	65	20	7	4	2,472
Geography	1	2	53	21	12	10	962
Geology	1	3	64	15	9	8	1,636
German	1	3	64	18	7	6	702
History	1	3	62	21	9	4	10,637
Literature	1	2	60	22	10	5	14,079
Mathematics	2	4	64	16	9	6	7,131
Music	3	2	53	24	12	8	2,503
Philosophy	1	3	65	18	8	5	1,570
Physics	1	3	67	12	9	7	3,907
Political Science	1	3	62	21	8	5	5,314
Psychology	2	4	65	18	7	5	17,578
Sociology	1	3	67	18	6	4	6,485
Spanish	2	3	60	20	10	6	1,739

partment within the institution receives a similar report based on scores reported to that department. A third set of distributions, based on all scores reported to all institutions, presents distributions for each intended major field group. A sample page from this third set, presented as Table 25, shows the frequency distributions for all 1975–76 examinees who intended to major in microbiology. In addition to graduate institution summary statistics reports, the GRE Program supplies the same kinds of information to undergraduate institutions. Each undergraduate institution report is based on the score information for examinees who identified that institution as the one in which they did their undergraduate work.

Although it was known for some time that GRE examinees are not restricted to seniors applying for admission to graduate schools, there was a general assumption that the proportion of examinees not included in this category was rather small. In 1970, Chaur C. Chen analyzed examinees' responses to background information questions for the Advanced Tests and found that the proportion of Advanced Test examinees classified as seniors was lower than expected, averaging about 60 to 70 percent. A similar study based on data from the following year (1970–71) was completed by Frances Swineford in 1971. Table 26 summarizes the examinee volume analysis in the report of the latter study. Of particular interest are the data for the Advanced Education, Engineering, Geography, and Music Tests.

In light of this information, questions were raised about the need to provide more clearly defined norms groups. The first percentile rank tables based on a normative group restricted to seniors plus nonenrolled college graduates who had no graduate school experience and who took the tests in 1974–75 were published as a supplement to the 1975–77 GRE *Guide*. The 1977–78 edition of the GRE *Guide* was the first to include this kind of information. These norms tables are probably more appropriate than the three-year aggregate norms for evaluating the performance of applicants for admission to graduate schools.

Descriptive Statistics for the Aptitude Test

In the preceding discussion of the development of the GRE scaled-score system, the fact that students selecting the various major fields show, on the average, different levels of developed abilities was well recognized as evidenced by its incorporation into the GRE scaled-score system. It is important that the magnitude of these differences be made known to score recipients who use GRE Aptitude Test scores as part of applicant information for making selection decisions. The descriptive data summarized in Table 27 are based on the same group described in the preceding paragraphs: seniors plus nonenrolled college graduates tested between 1974 and 1976. The table is based on examinee response to the background information question on undergraduate major field and shows Aptitude Test performance for each major field category. The major fields are grouped into the four main categories: humanities, social sciences, biological sciences, and physical sciences. Each of these is further divided into subgroups based on a logical structure: language versus nonlanguage majors in the humanities; education, history, business-commerce-communications in the social sciences; applied versus basic science in the biological sciences; and mathematics versus physical sciences in the physical sciences.

Table 27: Aptitude Test Performance of Seniors and Nonenrolled College Graduates Classified by Undergraduate Major Field

(Tested at National Administrations between October 1, 1974, and June 30, 1976)

HUMANITIES

Undergraduate Major Field	N	Verbal Ability		Quantitative Ability		Percent of Category
Subgroup A		Mean	Standard Deviation	Mean	Standard Deviation	
Linguistics	522	602	126	568	123	1.9
Other Foreign Languages	564	485	128	451	117	2.0
Classical Languages	638	635	107	556	110	2.3
English	18,883	577	112	499	115	67.8
French	2,731	560	107	505	105	9.8
German	1,051	573	106	533	116	3.8
Russian	482	607	116	555	116	1.7
Spanish	2,397	513	120	469	117	8.6
Far Eastern Languages	339	594	118	558	112	1.2
Near Eastern Languages	124	586	132	553	137	0.4
Italian	103	503	123	474	107	0.4
Subgroup B						
Archeology	295	562	118	519	117	0.9
Architecture	2,313	496	116	592	104	6.8
Fine Arts	4,379	499	112	471	113	12.8
Music	5,816	515	114	505	118	17.1
Philosophy	3,304	592	112	558	122	9.7
Religion	3,148	532	117	514	120	9.2
Speech	3,900	480	99	461	107	11.4
Art History	2,100	569	102	501	111	6.2
Comparative Literature	636	600	112	513	119	1.9
Dramatic Arts	2,210	534	109	487	118	6.5
Other Humanities	6,001	491	130	460	125	17.6
Total Subgroup A	27,834	570	116	502	116	44.9
Total Subgroup B	34,102	519	120	499	123	55.1
HUMANITIES	61,936	542	121	500	120	17.3

Table 27 (cont.)

SOCIAL SCIENCES

Undergraduate Major Field	N	Verbal Ability		Quantitative Ability		Percent of Category
Subgroup A		Mean	Standard Deviation	Mean	Standard Deviation	
Educational Administration	117	392	114	412	117	0.1
Education	22,027	458	103	460	113	27.8
Physical Education	3,632	420	93	460	111	4.6
Guidance and Counseling	295	457	112	434	119	0.4
Educational Psychology	212	477	115	492	118	0.3
Social Psychology	525	526	116	509	121	0.7
Psychology	37,866	530	106	523	116	47.6
Social Work	3,343	462	108	433	114	4.2
Sociology	11,280	489	115	469	122	14.3
Subgroup B						
American Studies	1,031	576	106	518	114	3.0
Slavic Studies	122	607	94	568	121	0.4
Geography	1,958	512	108	546	111	5.7
Anthropology	3,997	576	103	525	113	11.6
History	13,879	549	116	501	121	40.4
International Relations	1,277	558	110	534	119	3.7
Law	394	421	119	456	134	1.1
Government	11,700	526	113	506	121	34.1
Subgroup C						
Journalism	2,205	535	106	496	113	5.4
Business and Commerce	6,788	455	107	531	123	16.8
Communications	2,278	501	109	484	118	5.6
Economics	6,546	527	123	600	119	16.2
Industrial Relations	356	492	104	527	116	0.9
Library Science	560	482	111	442	107	1.4
Public Administration	618	477	111	489	119	1.5
Urban Development	946	517	111	530	124	2.3
Other Social Sciences	20,174	450	116	451	126	49.8
Total Subgroup A	79,097	496	112	491	120	51.4
Total Subgroup B	34,358	542	115	510	121	22.3
Total Subgroup C	40,471	474	119	496	135	26.3
SOCIAL SCIENCES	153,926	500	117	496	124	43.0

BIOLOGICAL SCIENCES

Undergraduate Major Field	N	Verbal Ability		Quantitative Ability		Percent of Category
Subgroup A		Mean	Standard Deviation	Mean	Standard Deviation	
Pharmacology	47	446	134	519	144	0.3
Audiology	516	480	89	476	102	3.4
Hospital Administration	256	444	97	476	127	1.7
Anatomy	50	510	124	567	139	0.3
Dentistry	213	465	98	535	113	1.4
Medicine	663	521	103	592	115	4.4
Nursing	6,337	504	96	478	101	42.2
Occupational Therapy	230	512	93	505	108	1.5
Optometry	18	497	146	593	126	0.1
Osteopathy	17	502	91	530	72	0.1
Pharmacy	687	478	111	590	93	4.6
Physical Therapy	465	499	94	530	100	3.1
Physiology	437	534	106	602	108	2.9
Public Health	413	466	97	490	121	2.8
Veterinary Medicine	522	492	96	557	101	3.5
Pathology	152	442	98	488	117	1.0
Nutrition	1,415	489	105	522	110	9.4
Home Economics	2,572	451	96	466	106	17.1
Subgroup B						
Genetics	263	570	102	648	92	0.6
Microbiology	1,904	522	100	588	97	4.1
Parasitology	32	451	131	466	115	0.1
Other Biological Sciences	10,847	481	111	528	121	23.2
Agriculture	2,545	452	107	541	106	5.5
Bacteriology	326	502	102	574	107	0.7
Biochemistry	1,809	563	108	655	98	3.9
Biology	22,189	528	105	580	108	47.6
Biophysics	144	576	104	676	102	0.3
Botany	1,053	543	105	591	102	2.3
Entomology	258	504	109	570	102	0.6
Forestry	959	504	96	599	93	2.1
Zoology	4,329	537	99	594	97	9.3
Total Subgroup A	15,010	489	101	500	113	24.3
Total Subgroup B	46,658	515	109	572	113	75.7
BIOLOGICAL SCIENCES	61,668	509	108	554	117	17.2

PHYSICAL SCIENCES

Undergraduate Major Field	N	Verbal Ability		Quantitative Ability		Percent of Category
Subgroup A		Mean	Standard Deviation	Mean	Standard Deviation	
Applied Mathematics	578	529	124	683	96	5.1
Statistics	209	488	139	664	102	1.8
Mathematics	8,591	525	121	671	101	75.7
Computer Science	1,975	530	120	673	99	17.4
Subgroup B						
Other Physical Sciences	4,022	449	121	594	119	11.3
Astronomy	216	601	104	683	101	0.6
Chemistry	7,891	529	115	647	100	22.2
Engineering, Aeronautical	534	515	105	677	85	1.5
Engineering, Chemical	1,719	482	130	678	91	4.8
Engineering, Civil	2,771	471	115	663	93	7.8
Engineering, Electrical	4,380	481	133	674	95	12.3
Engineering, Industrial	598	447	125	636	104	1.7
Engineering, Mechanical	2,349	462	128	668	93	6.6
Engineering, Other	2,077	503	113	672	92	5.9
Geology	3,889	532	105	606	100	11.0
Metallurgy	200	471	134	664	95	0.6
Mining	36	430	140	600	119	0.1
Oceanography	366	499	101	619	107	1.0
Physics	4,440	559	123	694	92	12.5
Total Subgroup A	11,353	526	121	672	100	24.4
Total Subgroup B	35,488	503	125	652	104	75.8
PHYSICAL SCIENCES	46,841	509	125	656	103	13.1

OTHER

Undergraduate Major Field	N	Verbal Ability		Quantitative Ability		Percent of Category
Subgroup		Mean	Standard Deviation	Mean	Standard Deviation	
Other	11,233	440	125	471	136	33.8
No Response	21,966	491	128	510	139	66.2
Total Subgroup	33,199	474	129	497	139	9.3
All Seniors and Nonenrolled College Graduates	357,570	508	120	528	133	100.0

Table 28 provides the same kinds of information based on responses to the background information question on intended graduate major field. The purpose of both tables is to point out the magnitude of the range of means for both verbal and quantitative ability scores and the relationships between Aptitude Test performance and educational experience and educational goals.

Because normative data on Aptitude Test performance by intended graduate major field are important in score interpretation for the purpose of selection, grouped score distributions by intended major field are now included in the GRE *Guide*. The score intervals used in these distributions are rather large, but the main statistical properties of the distributions can be observed.

When this technical manual went to press, the only normative information available for the new analytical measure was based on the data obtained in the first administration of the restructured Aptitude Test (October 1977). Although the examinee group of that administration is high scoring and includes a relatively high proportion of fellowship applicants, the relationships among the verbal, quantitative, and analytical scores for the four undergraduate major field categories may provide a useful guide for interpreting scores on the new measure. A brief summary of this information is given in Table 29.

Other Factors Interacting with Aptitude Test Performance

Other factors related to Aptitude Test performance have been examined for this same norms group. Three of these are sum-

Table 28: Aptitude Test Performance of Seniors and Nonenrolled College Graduates Classified by Intended Graduate Major Field

(Tested at National Administrations between
October 1, 1974, and June 30, 1976)

HUMANITIES

Intended Graduate Major Field Subgroup A	N	Verbal Ability Mean	Verbal Ability Standard Deviation	Quantitative Ability Mean	Quantitative Ability Standard Deviation	Percent of Category
Linguistics	970	586	133	553	127	6.5
Other Foreign Languages	333	483	136	448	126	2.2
Classical Languages	380	647	105	560	112	2.6
English	9,310	591	109	505	115	62.7
French	1,210	552	109	493	106	8.2
German	537	567	107	525	112	3.6
Russian	318	602	116	547	118	2.1
Spanish	1,206	493	121	450	117	8.1
Far Eastern Languages	336	595	109	547	115	2.3
Near Eastern Languages	171	602	114	548	127	1.2
Italian	66	502	129	445	100	0.4
Subgroup B						
Archeology	864	569	108	515	110	2.6
Architecture	3,048	512	117	588	107	9.3
Fine Arts	3,109	494	114	465	114	9.4
Music	4,923	515	115	507	119	14.9
Philosophy	2,005	603	109	567	122	6.1
Religion	5,205	527	113	524	123	15.8
Speech	3,230	477	100	457	107	9.8
Art History	1,809	565	107	490	109	5.5
Comparative Literature	793	610	111	521	112	2.4
Dramatic Arts	2,267	541	111	494	120	6.9
Other Humanities	5,692	497	127	467	124	17.3
Total Subgroup A	14,837	578	117	505	118	31.1
Total Subgroup B	32,945	522	119	504	124	68.9
HUMANITIES	47,782	539	121	504	122	13.4

SOCIAL SCIENCES

Intended Graduate Major Field Subgroup A	N	Verbal Ability Mean	Verbal Ability Standard Deviation	Quantitative Ability Mean	Quantitative Ability Standard Deviation	Percent of Category
Educational Administration	2,608	456	107	476	121	3.4
Education	22,420	471	107	472	116	29.4
Physical Education	2,558	426	95	470	115	3.4
Guidance and Counseling	6,968	478	107	466	116	9.1
Educational Psychology	2,447	507	106	500	114	3.2
Social Psychology	1,091	521	112	506	120	1.4
Psychology	26,429	534	108	525	118	34.7
Social Work	7,880	485	111	458	117	10.3
Sociology	3,759	495	121	479	128	4.9
Subgroup B						
American Studies	577	576	107	514	115	2.7
Slavic Studies	187	607	97	556	113	0.9
Geography	1,311	514	110	543	113	6.1
Anthropology	2,733	572	105	519	115	12.7
History	6,755	555	118	502	123	31.5
International Relations	2,910	552	114	524	121	13.6
Law	2,210	520	123	521	135	10.3
Government	4,770	526	120	503	126	22.2
Subgroup C						
Journalism	2,926	549	109	497	116	5.2
Business and Commerce	7,135	478	111	556	122	12.8
Communications	3,010	510	112	488	120	5.4
Economics	4,747	526	129	607	120	8.5
Industrial Relations	1,210	500	105	516	118	2.2
Library Science	6,687	551	112	481	112	12.0
Public Administration	5,569	498	110	493	121	10.0
Urban Development	3,693	525	111	535	124	6.6
Other Social Sciences	20,781	457	117	458	126	37.3
Total Subgroup A	76,160	496	113	490	121	49.7
Total Subgroup B	21,453	545	118	513	124	14.0
Total Subgroup C	55,758	494	120	500	130	36.3
SOCIAL SCIENCES	153,371	502	117	497	125	42.9

BIOLOGICAL SCIENCES

Intended Graduate Major Field Subgroup A	N	Verbal Ability Mean	Verbal Ability Standard Deviation	Quantitative Ability Mean	Quantitative Ability Standard Deviation	Percent of Category
Pharmacology	1,082	527	109	609	102	3.9
Audiology	1,067	488	95	483	104	3.8
Hospital Administration	2,193	485	104	517	119	7.9
Anatomy	633	508	98	556	105	2.3
Dentistry	474	477	101	560	117	1.7
Medicine	2,152	544	106	609	110	7.7
Nursing	6,130	507	97	481	101	22.0
Occupational Therapy	539	501	100	493	110	1.9
Optometry	100	497	95	585	107	0.4
Osteopathy	85	527	97	556	101	0.3
Pharmacy	568	462	111	571	107	2.0
Physical Therapy	1,715	481	98	526	103	6.1
Physiology	2,078	529	104	591	104	7.4
Public Health	2,621	512	112	529	120	9.4
Veterinary Medicine	2,476	523	100	587	98	8.9
Pathology	704	490	106	543	119	2.5
Nutrition	1,782	487	108	526	112	6.4
Home Economics	1,498	444	98	463	106	5.4
Subgroup B						
Genetics	1,161	557	101	616	97	3.4
Microbiology	3,379	513	105	574	106	9.8
Parasitology	150	510	114	544	109	0.4
Other Biological Sciences	13,819	500	112	549	121	40.0
Agriculture	1,993	455	112	541	109	5.8
Bacteriology	443	498	102	548	110	1.3
Biochemistry	3,027	547	109	640	101	8.8
Biology	4,509	521	112	574	114	13.0
Biophysics	337	565	115	664	101	1.0
Botany	1,379	551	103	588	101	4.0
Entomology	573	511	106	565	109	1.7
Forestry	1,041	509	97	592	101	3.0
Zoology	2,772	545	98	591	98	8.0
Total Subgroup A	27,897	504	105	533	118	44.6
Total Subgroup B	34,583	514	111	572	116	55.4
BIOLOGICAL SCIENCES	62,480	509	109	554	118	17.5

PHYSICAL SCIENCES

Intended Graduate Major Field Subgroup A	N	Verbal Ability Mean	Verbal Ability Standard Deviation	Quantitative Ability Mean	Quantitative Ability Standard Deviation	Percent of Category
Applied Mathematics	733	537	118	682	100	8.5
Statistics	616	514	118	679	101	7.1
Mathematics	3,394	527	130	676	105	39.2
Computer Science	3,922	523	128	669	100	45.3
Subgroup B						
Other Physical Sciences	4,258	460	124	602	121	13.9
Astronomy	423	595	106	689	95	1.4
Chemistry	4,765	524	118	650	100	15.6
Engineering, Aeronautical	460	507	110	672	91	1.5
Engineering, Chemical	1,604	479	130	672	96	5.2
Engineering, Civil	2,479	467	116	[illegible]	99	8.1
Engineering, Electrical	3,684	482	133	676	95	12.0
Engineering, Industrial	901	437	126	636	103	2.9
Engineering, Mechanical	1,737	461	128	666	93	5.7
Engineering, Other	2,396	505	115	667	98	7.8
Geology	3,615	530	104	604	102	11.8
Metallurgy	220	479	136	665	95	0.7
Mining	45	493	112	615	111	0.1
Oceanography	1,040	521	104	620	100	3.4
Physics	3,011	562	125	702	87	9.8
Total Subgroup A	8,665	525	127	674	102	22.0
Total Subgroup B	30,638	500	126	649	106	78.0
PHYSICAL SCIENCES	39,303	506	126	655	105	11.0

OTHER

Intended Graduate Major Field Subgroup	N	Verbal Ability Mean	Verbal Ability Standard Deviation	Quantitative Ability Mean	Quantitative Ability Standard Deviation	Percent of Category
Other	10,526	449	123	470	133	19.3
Undecided	10,487	521	123	537	132	19.2
No Response	33,621	500	127	521	137	61.5
Total Subgroup	54,634	494	128	514	137	15.3
All Seniors and Nonenrolled College Graduates	357,570	508	120	528	133	100.0

Table 29: Summary Statistics for the Aptitude Test Performance of Seniors and Nonenrolled College Graduates Tested in October 1977, Classified by Undergraduate Major Field

Undergraduate Major Field	N	Verbal Ability		Quantitative Ability		Analytical Ability	
		Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Humanities	4,905	560	122	522	115	546	114
Social Sciences	13,180	512	118	516	117	527	120
Biological Sciences	6,493	515	108	563	107	544	110
Physical Sciences	4,952	537	121	673	107	587	114
TOTAL GROUP*	33,059	523	119	553	127	542	119

*The total group includes 3,529 examinees who did not indicate major field.

Table 30: Aptitude Test Performance of Seniors and Nonenrolled College Graduates, Classified by Graduate Degree Objective

Graduate Degree Objective	N	Verbal Ability		Quantitative Ability		Percent of Total Group
		Mean	Standard Deviation	Mean	Standard Deviation	
Nondegree study	4,814	471	134	511	142	1.33
Master's degree (M.A., M.S., M.Ed., etc.)	196,923	487	114	507	129	55.07
Intermediate (such as Specialist)	7,496	494	114	515	127	2.10
Doctorate (Ph.D., Ed.D., etc.)	111,333	544	116	561	130	31.14
Postdoctoral study	9,721	569	114	593	128	2.72
No response	27,283	497	131	523	140	7.63
TOTAL GROUP	357,570	508	120	527	133	100

Table 31: Aptitude Test Performance of Seniors and Nonenrolled College Graduates, Classified by Citizenship and by Primary Language

Classification	N	Verbal Ability		Quantitative Ability		Percent of Total Group
		Mean	Standard Deviation	Mean	Standard Deviation	
CITIZENSHIP						
American	316,693	516	115	528	132	88.57
Other	21,534	400	132	542	142	6.02
No Response	19,343	495	127	515	138	5.41
PRIMARY LANGUAGE						
English	314,531	514	116	529	132	87.96
Other	21,881	427	134	523	142	6.12
No Response	21,158	494	127	514	138	5.92
TOTAL GROUP	357,570	508	120	527	133	100

marized in Tables 30 and 31: graduate degree objectives, citizenship, and primary language. These tables are less useful than those described in the preceding paragraphs because they consider each factor independently of the others, even though there is likely to be a complex interaction among all factors. Of interest is the proportion of examinees in certain categories. For example, more than 6 percent are not American citizens, and more than 6 percent have indicated that they communicate best in a language other than English.

Additional information describing the GRE population tested during the academic year 1975–76 is provided in a GRE report by Robert A. Altman and Paul W. Holland, *A Summary of Data Collected from Graduate Record Examinations Test-Takers During 1975–76*, Data Summary Report #1, March 1977, Educational Testing Service.

References

Chen, C. C. *Graduate Record Examinations Advanced Tests: Summary of responses to background information questions, October 1969 to July 1970 administrations* (Statistical Report SR-70-99). Princeton, N.J.: Educational Testing Service, 1970.

Dressel, P. L. Some remarks on the Kuder-Richardson reliability coefficient. *Psychometrika*, 1940, *5*, 305–310.

Kuder, G. F., & Richardson, M. W. The theory of the estimation of test reliability. *Psychometrika*, 1937, *2*, 151–160.

Levine, R. *Equating the score scales of alternate forms administered to samples of different abilities* (Research Bulletin RB-55-23). Princeton, N.J.: Educational Testing Service, 1955. [Submitted as a doctoral thesis, Syracuse University, 1955.]

Schultz, M. K., & Angoff, W. H. The development of new scales for the Aptitude and Advanced Tests of the Graduate Record Examinations. *The Journal of Educational Psychology*, 1956, *47*, 285–294.

Swineford, F. *Graduate Record Examinations Advanced Tests: Summary of responses to background information questions, 1970–71 administrations* (Statistical Report SR-71-108). Princeton, N.J.: Educational Testing Service, 1974.

Swineford, F. *An assessment of the Kuder-Richardson formula (20) reliability estimate for moderately speeded tests.* Unpublished report, Educational Testing Service, 1973.

Wallmark, M. M. *A rescaling study of the Graduate Record Examinations Advanced Tests* (Statistical Report SR-69-4). Princeton, N.J.: Educational Testing Service, 1969.

Chapter 6

VALIDITY OF THE GRADUATE RECORD EXAMINATIONS*

Of the various characteristics of a test, its validity is often the focus of greatest interest. Yet the term is ambiguous; it can be interpreted in a variety of ways. Defined simply, validity is the degree to which a test reflects the truth about the characteristics of the person whose traits it purports to measure. Thus, validity is as much concerned with what a test claims to measure as with the means by which it measures.

The GRE Aptitude Test is defined as a measure of developed verbal, quantitative, and analytical abilities. These abilities are scholastic in nature and broadly applicable. They are assumed to be the product of the interaction of personal characteristics and experience and to be related to achievement in activities requiring those skills. The Advanced Tests are designed to measure an individual's mastery of a given discipline, which may be defined rather specifically in terms of the typical undergraduate curriculum or the usual expectations of graduate students in a field. What a test does *not* measure may also be usefully identified in assessing validity. For example, the Aptitude Test cannot be said to be valid for measuring creativity or "raw" intelligence because it is not designed to measure these traits.

It could rightly be said that the GRE Advanced History Test, which covers American and European history, is not a valid measure of a student's knowledge of ancient Chinese history. Even if all students who do well on the Advanced History Test also score high on a test of ancient Chinese history, the Advanced History Test does not purport to measure that particular domain of knowledge and cannot be valid for that purpose. Validity must be judged in the context of the purposes of a test.

The ways of assessing validity are generally expressed as different kinds of validity. Because these kinds of validity are really ways of articulating questions about how well a test measures up to its claims rather than clearly defined aspects of a concept, they cannot be viewed as distinct and unrelated. For example, "content" and "construct" validity, though differentiated as terms, may not always be extricable in situations in which validity is being explored.

The kinds of validity most frequently referred to are content validity, construct validity, and criterion-related validity. The definitions of these terms quoted in the following paragraphs are taken from *Standards for Educational and Psychological Tests* (American Psychological Association, 1974).

"Evidence of content validity is required when the test user wishes to estimate how an individual performs in the universe of situations the test is intended to represent. Content validity is most commonly evaluated for tests of skill or knowledge; it may also be appropriate to inquire into the content validity of personality inventories, behavior checklists, or measures of various aptitudes" (p. 28). Content validity has special relevance to the Advanced Tests since these examinations must represent subject fields accurately and produce appraisals of knowledge that are fair regardless of the fact that undergraduate curriculums vary from institution to institution. However, the question of content validity is also appropriate to an evaluation of the Aptitude Test.

Construct validity concerns the relevance and legitimacy of the skill domains being tested. "Evidence of construct validity is not found in a single study; rather, judgments of construct validity are based upon an accumulation of research results. In obtaining the information needed to establish construct validity, the investigator begins by formulating hypotheses about the characteristics of those who have high scores on the test in contrast to those who have low scores. Taken together, such hypotheses form at least a tentative theory about the nature of the construct the test is believed to be measuring" (p. 30). In considerable part, the construct validity of the GRE rests upon decades of psychometric research, indicating the sorts of ability that play a critical role in most types of intellectual work, and upon even more extensive educational experience, indicating that frequently the best predictor of future success in an academic field is early competence revealed by a subject-matter test. Construct validation requires constant attention, however, to ensure that a test is actually measuring the construct intended. For example, a reading comprehension test should not be so complicated in content as to stress reasoning instead of reading, or a mathematics test should not use language that places a premium upon knowledge of vocabulary.

"Criterion-related validities apply when one wishes to infer from a test score an individual's most probable standing on some other variable called a criterion. Statements of predictive validity [for example] indicate the extent to which an individual's future level on the criterion can be predicted from a knowledge of prior test performance. . . . For many test uses, such as for selection decisions, . . . predictive validity provides the appropriate model for evaluating the use of a test or test battery" (p. 26). Predictive validity is particularly important to the GRE Program because the examinations are used to select students likely to succeed in graduate study, but validity based on other criteria (such as self-reported undergraduate grades) has also been explored.

Content Validity

Concern for the content validity of the Aptitude Test is reflected in test specifications based on: 1) diversity of topics and points tested; 2) coverage of fundamental concepts and skills; and 3) use of a variety of methods of testing skills—for example, antonyms, analogies, sentence completions, and reading comprehension in the verbal measure and computation problems, data interpretation, and quantitative reasoning questions in the quantitative measure. The question of content validity is fundamentally whether the test adequately samples the domains of verbal, quantitative, and analytical skills.

The first and most important step taken to help assure the content validity of the Advanced Tests is the direct involvement of scholars and teachers in the discipline of each test in writing, reviewing, revising, selecting, and approving questions for that test. College professors who are actively engaged in teaching at recognized institutions and who are therefore believed to be fa-

---

*Parts of this chapter have been excerpted from *Validity and the Graduate Record Examinations Program* (Willingham, 1976) and "Predicting Success in Graduate Education" (Willingham, 1974).

miliar with the content of typical undergraduate curriculums and the requirements of graduate study in their disciplines serve as members of the committees of examiners.

Several additional steps are taken to aid the examiners in striving for content validity. Probably the most important additional step is the systematic and continual feedback to the examiners of performance data of examinees on test questions. In addition, questions about the educational backgrounds and goals of examinees are periodically included in the test book. Examinees respond to these questions just prior to taking the tests. The responses to the questions are analyzed to show the test performance of student groups with different backgrounds and goals, but the responses have no influence on reported scores. Occasionally, a more extensive questionnaire on student backgrounds and reactions to the test is given to samples of examinees at a test administration. Responses to the questionnaire are mailed back by the examinees following the administration.

From time to time, a representative panel of college professors reviews a test's specifications and actual test copies in considerable detail. Some tests are routinely reviewed before printing by professors not on the committee of examiners. Inspection copies of tests may be requested by college presidents, deans, or graduate department chairmen; forms to be used in evaluating the tests are routinely sent with inspection copies and are completed by faculty members who review them. Articles about certain Advanced Tests appear occasionally in appropriate professional journals, and presentations on Advanced Tests are made sometimes at professional meetings. These articles and presentations help secure feedback from test users about the content validity of the tests.

Construct Validity

Construct validity concerns the degree to which the domains of skills tapped by the test appear to be related to those domains as defined in other contexts. Construct validity of the Aptitude Test is evinced by the tendency for people in fields requiring quantitative skills to have relatively high quantitative scores, and for people in fields requiring verbal skills to have relatively high verbal scores. For example, two of the highest correlations between an Advanced Test and the quantitative ability measure are for economics and mathematics, both fields in which quantitative skills are important. High correlations between the verbal ability scores on the Aptitude Test and scores on the Advanced Literature in English Test, the Advanced Education Test, and the Advanced Philosophy Test suggest that the Aptitude Test is truly measuring a verbal construct underlying performance in those fields. (See Table 13 on page 31.)

The results of predictive validity studies also suggest that verbal and quantitative ability constructs are appropriately reflected in the Aptitude Test. In those scientific fields where quantitative ability counts most, the GRE Aptitude Test quantitative ability score is typically a better predictor than the verbal ability score. Correspondingly, the GRE Aptitude Test verbal score tends to be more valid in such verbally oriented disciplines as English and education than in scientific fields. Intercorrelations of the verbal and quantitative measures are sufficiently low (.50–.60) to suggest the independence of the constructs.

One of the most common ways of investigating construct validity is through factor analysis. Examination of the relationships among questions in a test contributes to an understanding of the abilities that affect performance and has implications for test development.

The decision of the GRE Board to offer a restructured Aptitude Test in October 1977 was based on the presupposition that the restructured test should measure the same verbal and quantitative constructs as those measured before and that a new measure should tap a construct with unique dimensions. Factor analyses were performed to determine whether projected changes in the verbal and quantitative measures would be appropriate to the constructs as defined by the original verbal and quantitative measures and whether the analytical question types under study would have qualities separating them from the verbal and quantitative domains.

In the first factor analysis (Powers, Swinton, and Carlson, 1977) undertaken in relation to investigating the possibility of restructuring the Aptitude Test, principal factor solutions were computed for the responses of two random samples, each consisting of 8,000 examinees, taking one of two forms of the GRE Aptitude Test. In addition, the factors of the test forms were extended into each of eight experimental subtests, which were administered along with the final forms in a spiral design. These experimental subtests contained variations of the Aptitude Test content considered as potential constituents of a restructured Aptitude Test. Four were verbal and four were quantitative in nature. For example, one subtest contained short reading comprehension passages considered for potential inclusion, as well as longer ones that had been used exclusively in the original operational Aptitude Test. Others contained reading passages with homogeneous content—either science passages *or* humanities and social studies passages. Still others contained only data interpretation questions, used sparingly in the original test, and quantitative comparisons, a new mathematics question type not then used in the Aptitude Test. (For an extended discussion of Aptitude Test content, see Chapter 3.)

Both forms revealed three factors reflecting the global structure of skills assessed by the Aptitude Test. These three major dimensions of question covariance accounted for about three-fourths of the common variance in each form. Factor I was identified as a *general quantitative* factor, accounting for nearly 30 percent of the common variance in each form. Most of the quantitative questions but none of the verbal questions loaded highest on this factor.

Factor II, accounting for about one-fourth of the common variance in each form, was identified as a *reading comprehension* (connected discourse) factor. Almost all the questions associated with reading passages exhibited their highest loadings on this factor, and most of the sentence completion questions showed substantial loadings on this factor. Physical science passages appeared less strongly related to the comprehension dimension than passages based on humanities or social science content.

Factor III was identified as a *vocabulary* factor (words and concepts in isolation) and accounted for about 20 percent of the common variance of each form. Approximately 90 percent of the antonym questions and 70 percent of the analogy questions loaded highest on this factor.

The other less important factors (in terms of variance explained) identified for the first form were as follows: Factor IV, contributing 5 to 10 percent of the common variance, was identified as an *elementary algebra* factor since each of the five questions having their highest loading on the factor involved algebraic notation. Factor V, also accounting for 5 to 10 percent of the common variance, appeared to represent *speed of response to discrete verbal* questions (as opposed to speed in questions associated with reading passages). Factor VI, accounting for less than 5 percent of the common variance, was a dimension of variance underlying certain data interpretation questions from the quantitative section. These questions were based on a complex graph or table and required the extraction of information from that table or graph. Thus, this factor was identified as the *ability to extract information.* Factor VII, also accounting for less than 5 percent of the common variance, was designated an *applications: word problems* factor. Factor VIII, accounting for less than 5 percent of the common variance, seemed to represent a factor of *reading speed in comprehension passages.* From the small percent of variance accounted for by this factor, it was concluded that speed does not play an important role in the reading comprehension section of the test.

The results of the factor analysis of the second form resembled the results of the first. However, two differentiated data interpretation factors, comparable to Factor VI in the first test form analyzed, were discovered and called *data interpretation: extraction and manipulation* and *data interpretation: extraction.* In addition, a factor termed *reading comprehension: scientific/technical* was identified; it accounted for less than 5 percent of the test's common variance. Two other minor factors not identified in the first form were found: *quantitative speed* (accounting for less than 5 percent of the variance) and what was dubbed *easy antonyms,* though interpretation of this latter factor was somewhat problematic.

In the second form of the test, two factors isolated in the first form were not found: *elementary algebra* (characterized by the presence of algebraic notation) and *word problems.* The two forms were, however, very closely similar since the first three rotated factors, which together account for approximately three-fourths of the *common* and 40 percent of the *total* variance in each form, represent the global skills tapped by the GRE Aptitude Test before restructuring—one quantitative and two verbal factors. The quantitative factor is general in nature by virtue of its high loadings on most of the quantitative items. The two verbal factors define abilities to deal with connected discourse (reading comprehension passages and sentence completions) and with words in isolation (antonyms and analogies). None of the remaining factors explains more than 10 percent (and most less than 5 percent) of the common variance.

The factor analysis not only identified the construct structure of the verbal and quantitative measures but also resulted in recommendations concerning proposed alterations in the test. The proposed inclusion of short as well as long passages in a restructured Aptitude Test was supported because the relationships among the questions associated with the shorter experimental passages were as well explained as those associated with the longer passages used in the operational forms.

Since the content of experimental subtests containing only scientific passages was not as well explained by the operational test factors as subtests containing nonscientific passages, the proposed provision of separate reading options (one with scientific, the other with humanities/social studies content) was abandoned. Such a change would have made the test a different experience for science students than for the other students.

More than 80 percent of the established common variance of each of the experimental quantitative pretests was explained by the factors found in the operational test. Thus a change to include quantitative comparison questions in a restructured test was considered acceptable in view of the objective to retain the construct validity of the original quantitative measure.

From the above discussion it can be observed that three decisions concerning test specifications for the restructured test were related to construct validity as investigated by factor analysis: 1) the decision to include short as well as long passages in the reading comprehension part of the test; 2) the decision *not* to separate reading comprehension passages by subject matter content into two optional modules; and 3) the decision to include some quantitative comparison questions in the quantitative ability measure to reduce the testing time, because these questions had high loadings on the general quantitative factor.

As stated earlier, evidence of construct validity is not found in a single study; rather, judgments of construct validity are based on an accumulation of research results. This accumulation is usually based not only on several studies using similar methods, but also on investigations using a variety of different techniques. Campbell (1960) and Campbell and Fiske (1959) have pointed out that to demonstrate construct validity it is necessary to show not only that the measure correlates highly with certain other variables, a process referred to as convergent validation, but also that it does not correlate significantly with certain other variables from which it should differ, a process referred to as discriminant validation. A good deal of research has been directed at establishing these two types of construct validity for the new analytical measure.

Evidence of the discriminant validity of the new types of questions has been obtained through a variety of judgments made by faculty, examinees, and others about whether the questions seem to measure something different from the verbal and quantitative skills also assessed by the Aptitude Test. Correlational analyses have shown that experimental tests containing the new analytical questions are in general slightly more related to the verbal and/or quantitative portions of the test than the verbal and quantitative portions are to each other.

Each of the three types of questions included in the new analytical ability measure of the Aptitude Test introduced in October 1977 exhibits, in differing degrees, variance not shared with the verbal or quantitative measures. Results of factor analysis studies conducted to date suggest that the logical diagrams questions have somewhat more in common with the quantitative measure than with the verbal section. A second type of question, analytical reasoning, shows this same pattern. The third type, analysis of explanations, however, has slightly more in common with the verbal section than with the quantitative section of the test. After statistical removal of the verbal and quantitative factors, however, there remains a unique interpretable dimension for each of the three types of analytical questions. Thus, results of factor analysis suggested that the addition of an analytical measure to the Aptitude Test would be supplementary rather than redundant.

Criterion-Related Validity

Because the purpose of the Graduate Record Examinations is to select applicants who will be best prepared to succeed in a graduate study program, the actual relationship between performance on the GRE and performance in a graduate program is an important concern.

Predictive Validity

Of the various ways of exploring criterion-related test validity, determining the predictive effectiveness of the examination is the most practical. Correlational analysis has been the principal research design for evaluating methods of selection, particularly in the case of easily quantifiable selectors such as test scores and undergraduate grades. One or more *predictors* (measures of student potential used for selection) may be evaluated by the extent to which they accurately forecast one or more *criteria* (measures of student success). The value of a predictor for selecting students is usually considered to vary directly with the size of its correlation with the criterion (Cronbach, 1971). This correlation, the validity coefficient, ranges from a chance relationship of 0.00 to a perfect relationship of 1.00, though negative coefficients can occur and perfect validity is not closely approached in practice. Usually more than one predictor is involved (for example, a test and a grade average), and in some cases a statistically weighted composite of the predictors is typically more useful for selection purposes than either predictor alone.

Although correlational analysis has conceptual simplicity, its application to the study of graduate student selection is complicated by a number of serious and often insoluble problems. First, in a plan to study the effectiveness of predictors, a decision must be made concerning the criteria by which to judge their effectiveness. Graduate grades are readily available and relevant indications of success, but many faculty members doubt that even reliable grades represent the most important outcomes of education. Comprehensive examinations are limited. Faculty ratings tend to be unreliable. Whether a student attains the Ph.D. depends on academic persistence and probably does not differentiate very well the most promising scholars and professionals. Yet waiting for proof of scholarship and contributions to a discipline could result in indefinite postponement of a predictive validity study. No criterion will be totally satisfactory, and the use of several relevant criteria must inevitably represent a compromise.

Second, almost any information on graduate student performance accumulates slowly. First-year grades are generally the earliest obtainable information for use as a criterion. Information on Ph.D. completion may not be collectible for several years after the predictor information has been recorded. For example, the analytical ability measure of the GRE Aptitude Test, introduced in October 1977, will not be subjected to predictive validity studies until criterion information is available—the end of 1979 at the very earliest; studies using the criterion of Ph.D. completion will not be possible until the early to middle 1980s.

Another serious problem is that, when studies focus on particular institutions and departments, as they must, the number of students may be so small that findings will fail to be statistically significant. Repeated studies on small groups may result in widely varied findings, with a predictor appearing very effective one year and ineffective another. Appendix II presents, for the GRE Advanced Tests, summaries of predictive validity studies relevant to each test for which they are available. These studies illustrate the problem of numbers, especially in relatively small fields. Concern for content and construct validity is especially important when correlational studies are difficult to interpret because of the small number of students involved.

An equally serious problem that results in deflated validity coefficients is the restricted range of students' performance on both predictors and criteria. From the standpoint of research design, an ideal method of studying predictive validity would be to collect selection information for all applicants, admit a random sample of applicants, and then examine the relationship between the criterion (performance in graduate school) and the predictors. However, to admit students on a random basis is probably neither practical nor ethical. Graduate applicants generally represent a highly select group with respect to academic ability and past performance. By the time students are admitted to departments, further restriction of range is introduced either directly (when the GRE and undergraduate grades have been used in selection) or indirectly (when other related variables have been used). Restriction in range on one or more of the predictor variables under consideration makes it difficult to obtain a clear assessment of the actual value of the predictors involved since observed validity coefficients tend to be lower than would be the case if a full range of talent (a group representative of all applicants) could be included in departmental samples.

In recent years, GRE verbal ability scores for examinees nationally have had standard deviations of approximately 125, and the standard deviations of GRE quantitative ability scores have been approximately 135. In departmental samples such as those involved in many validity studies, standard deviations of 75 to 90 on one or both of these variables are not uncommon, indicating that the range of ability available for study is considerably less than that in the total group of individuals taking the GRE nationally.

Restriction in the range of criterion values also complicates the interpretational outlook. If criterion values such as graduate grades vary only over a very limited range (A to B), differences in student performance may not be measured reliably. This tends to lead to underestimation of the overall utility of a predictor (Wilson, 1977).

The effect of restriction of range is often seen in studies in which a number of possible predictors are examined—only some of which are actually used. Predictors not in use would be expected to show a spuriously high correlation with the criteria because students are more heterogeneous in those respects; the range of distinctions has not been restricted by previous selection.

Although a predictive validity study is intended to check on the effectiveness of predictors, it is not intended to identify from an infinite list of possible predictors those that should be used because they provide high correlations. Predictors and criteria should have reasonable reliability (stability and freedom from distortion), educational relevance, and acceptability in terms of economic and administrative feasibility and ethical considerations. Faculty recommendations and ratings, though educationally relevant, are frequently unreliable; that is, the assessments of various judges are not comparable. On the other hand, predictors with the highest validity coefficients may not be educationally relevant. For example, the variable most highly related to the academic success of students might theoretically be possession of an automobile, but this variable is clearly neither an educationally sound nor an ethically acceptable predictor.

The correlations obtained in predictive validity studies, if interpreted in light of the limitations of such studies and with reference to educational and social values, are valuable information. A number of institutions and agencies have participated in predictive validity studies including the GRE as predictors, and results of some of those studies are summarized in Table 32. The studies summarized here include all or some of the following predictors: GRE Aptitude Test verbal and quantitative ability scores, GRE Advanced Test scores in students' chosen fields (thus, the content varies, depending on the department concerned), a GRE composite (usually the average of two or three GRE scores, though this composite was occasionally weighted statistically), under-

Table 32: Median Validity Coefficients† for Various Predictors and Criteria of Success in Graduate School

Predictors	Criteria of Success				
	Graduate GPA	Overall Faculty Rating	Department Examination	Attain Ph.D.	Time to Ph.D.
GRE verbal score	.24 46	.31 27	.42 5	.18 47	.16 18
GRE quantitative score	.23 43	.27 25	.27 5	.26 47	.25 18
GRE Advanced Test score	.30 25	.30 8	.48 2	.35 40	.34 18
GRE Composite	.33 30	.41 8	*	.31 33	.35 18
Undergraduate GPA	.31 26	.37 15	*	.14 30	.23 9
Recommendations	*	*	*	.18 15	.23 9
GRE-GPA Composite (weighted)	.45 24	*	*	.40 16	.40 9

†The lower number in each pair (set in smaller type) represents the number of coefficients upon which each median is based. See pages 60-62 for a list of the validity studies summarized here.

*No data available.

graduate grade-point average (undoubtedly computed in different studies in different ways that were seldom specified very carefully), recommendations [almost exclusively from three extensive studies of National Science Foundation fellowship applicants by Creager (1965) and Rock (1972), where the average rating of several letters of reference was used], and a weighted composite of GRE and grade-point average. The criteria of success in graduate school include graduate grade-point average, faculty ratings (typically representing the composite judgments of several faculty members concerning professional promise or overall success as a graduate student), departmental examinations (very few cases), and attainment of the Ph.D. This last factor typically means attaining the degree within a certain number of years, so a time element is also involved and has been formalized in the time-to-Ph.D. criterion by assigning criterion scores to students according to years elapsed between B.A. and Ph.D. All the data concerning Ph.D. attainment come from two studies by Creager (1965).

The studies summarized in Table 32 cover the period from 1952–1972, although about half were dated during the last five years of the span (see the list of studies on pages 60 to 62). Half of these studies were published; the other half were institutional reports or theses. Many of the studies were earlier described individually by Lannholm (1968, 1972). Since a report of more recent studies, as well as current studies sponsored by the GRE Board, has not yet been published, studies completed since 1972 are not summarized here.

The 43 studies in this summary include 138 independent sets of data, usually corresponding to departments, though occasionally representing some broader group such as first-year students across several departments. Individual sets of data are based on 20 to 1,479 students (median N = 80). The total number of students included in all studies is 21,214, and the total number of validity coefficients is 616.

Predictability of Graduate Success. The studies represented in Table 32 vary widely in quality and scope. Some are based on small samples, making individual correlations unreliable. But the medians based on more than just a few coefficients should give a dependable idea of how valid these predictors are and how predictable are the various criteria of graduate success. Insofar as possible, the same data have been sorted by major field and presented in Table 33 to illustrate differential validity of the predictors for different disciplines. Several observations can be made from these tables.

- Validity coefficients for the various predictors and composites (against the graduate grade-point average criterion) tend to be about .15 lower than corresponding median coefficients at the undergraduate level (Fishman and Pasanella, 1960).
- The undergraduate grade-point average is a moderately good predictor of graduate grade-point average and faculty ratings; it is a poor predictor of whether a student will attain the Ph.D. Depending on the success criterion used, the GRE composite is

Table 33: Median Validity Coefficients† for Five Predictors of Graduate Success (Variously Defined*) in Nine Fields

Predictors	Biological Science	Chemistry	Education	Engineering and Applied Science	English	Math	Physics	Psychology	Social Science
GRE verbal score	.18 7	.22 14	.36 15	.29 11	.21 6	.30 6	.02 6	.19 23	.32 11
GRE quantitative score	.27 8	.28 13	.28 14	.31 10	.06 6	.27 6	.21 6	.23 22	.32 10
GRE Advanced Test score	.26 5	.39 9	.24 6	.44 7	.43 3	.44 5	.38 5	.24 17	.46 5
Undergraduate GPA	.13 2	.27 7	.30 5	.18 4	.22 4	.19 4	.31 4	.16 15	.37 6
GRE-GPA Composite (weighted)	.35 3	.42 6	.42 7	.47 4	.56 2	.41 3	.45 2	.32 4	.40 5

†The lower number in each pair (set in smaller type) represents the number of coefficients upon which each median is based. See pages 60-62 for a list of the validity studies summarized here.

*In sets of data where two criteria were included, one was selected for the purposes of this table in the following order of priority: GPA, Attain Ph.D., Department Test, and Faculty Rating.

either slightly more valid or substantially more valid than the undergraduate grade-point average.

- The GRE quantitative ability score is typically a better predictor in those scientific fields where quantitative ability counts most. The reversal in the field of mathematics may be due to restriction in the range of quantitative ability scores because of heavy emphasis on this variable in selection. Correspondingly, the GRE verbal ability score tends to be more valid in verbally oriented disciplines such as English and education. Otherwise the pattern of validity coefficients is fairly similar from one discipline to the next.
- The GRE Advanced Test score is evidently the most generally valid predictor among those included. In seven of the nine disciplines in Table 33, it has the highest validity among the three GRE scores. In eight of the nine fields, it has higher validity than undergraduate grade-point average.
- Recommendations appear to be a fairly poor predictor of whether a student will successfully complete a doctoral program.
- The comprehensive departmental examination seems a somewhat more predictable criterion than the others examined here. This is an uncertain conclusion because the available data are sparse, but the conclusion is consistent with the reasonable assumption that such a criterion should be more reliable than the others represented.
- A weighted composite including undergraduate grade-point average and one or more GRE scores typically provides a validity coefficient in the .40–.45 range for various criteria of success and for different academic fields. This is somewhat higher than the validity of GRE scores alone. The composite of undergraduate grade-point average and GRE provides a substantially more accurate prediction than does undergraduate grade-point average alone. This is the case for each success criterion and practically every academic discipline.

Utility of Current Predictions. What overall evaluation can be made of the extent to which success in graduate school is predictable? Brogden (1946) was the first to demonstrate that the correlation coefficient " is the ratio of the increase obtained by selecting above a given standard score on the test to the increase that would be obtained by selecting above the same standard score on the criterion itself" (p. 68). Or, as Cronbach (1971) later stated from the standpoint of utility theory, the correlation coefficient "expresses the benefit from testing as a percentage of the benefit one could get from perfect prediction of outcomes" (p. 496). Thus Table 32 indicates that in most fields the value of prediction by the GRE and grade-point average composite amounts to about 40 percent of the benefit that could accrue if prediction were perfect.

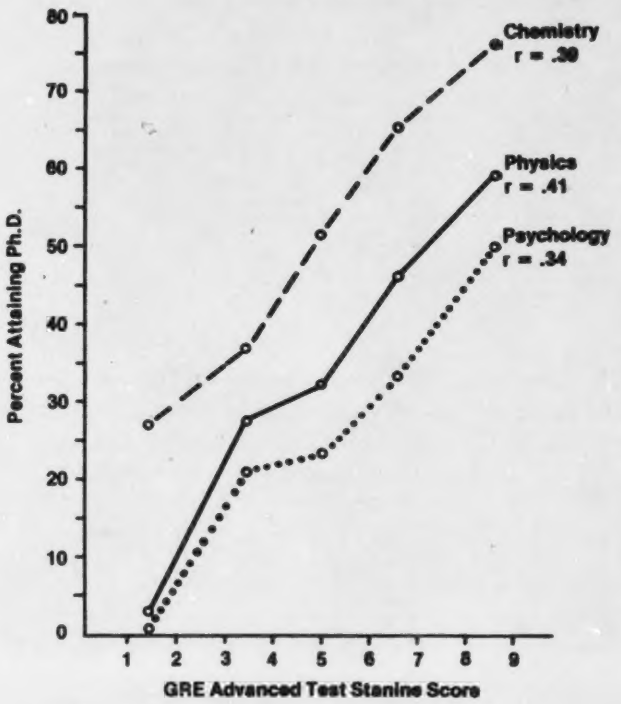
These validity coefficients, in fact, underestimate the usefulness of prediction in graduate admissions because they are based upon students actually admitted rather than the full range of students who apply. There are accepted procedures for correcting this restriction in range. The resulting validity coefficients are always higher, and may be substantially so if only a small proportion of applicants are accepted. For example, if a department selects only those students above the mean of its applicants on whatever admissions criterion it uses and the validity of that measure is .40 in the admitted group, the validity would be .59 if all applicants were admitted. In actual practice, corrections for restriction are usually not routinely made because the reasonableness of underlying assumptions and, therefore, the accuracy of corrections are difficult to ascertain.

How useful validity is in practical terms depends also on the cost of gathering the predictor information, the proportion of students selected, and the importance of the decision. A small correlation can produce a large benefit if the proportion of students selected is low. Finally, a given validity coefficient will have more practical value if the selection decision is important, and the selection decision is more important if it is irreversible.

A validity coefficient of .20 might be described as modest and one of .40 as moderate. The conditions of graduate student selection are generally favorable to using predictors of even modest validity. In many departments only a small proportion of students are accepted; the decisions are quite important to the student and to society; and the decisions are typically irreversible. There seems little doubt that the GRE and the undergraduate grade-point average are providing quite useful information in most situations, particularly since a given correlation represents greater usefulness in selecting among the applicant population than its size indicates, if the study has focused on admitted students only.

Figure 6 illustrates graphically the level of benefit likely to accrue from using predictors that are valid to the extent indicated. Students at high ability levels are far more likely to attain the Ph.D. than those at low levels. The figure also illustrates that many

Figure 6: Usefulness of GRE Advanced Test Scores for Predicting Ph.D. Attainment in Three Fields (Creager, 1965)



students fail to attain the degree, even among talented NSF fellowship applicants. And in these samples reported by Creager (1965), there are substantial differences in attainment rates among fields.

It should be emphasized also that validity studies at particular schools and departments give varying results. Such variability is exacerbated by the small samples often used, but real variations do occur. For this reason, the GRE Board encourages local studies to enable institutions to justify selection procedures and utilize available information to maximum benefit.

Other Evidence of Criterion-Related Validity

Although prediction is clearly of greatest concern in establishing the validity of the Graduate Record Examinations, the criterion of undergraduate grade-point average (as reported by students on GRE answer sheets) permits study of large numbers of students. In research on restructuring the GRE Aptitude Test, self-reported undergraduate grade-point average was used as a means of evaluating the potential usefulness of various measures of analytical ability. The correlations for the experimental analytical types of questions, three of which are components of the analytical measure introduced in October 1977, are reported on page 23. At the same time, correlations between verbal and quantitative ability scores and undergraduate grades were obtained. Tables 34 and 35 show the intercorrelations of verbal and quantitative measures in the October 1975 and December 1975 administrations of the GRE Aptitude Test. Since correlations were obtained for several spaced samples of various groups, the ranges of the coefficients are provided, along with ranges of scaled score means and standard deviations. The samples are slightly higher-scoring than representative samples because students answering background questions (on major field, etc.) tend to be higher-scoring on the average than the total population.

Table 34: Correlations of Verbal and Quantitative Ability Scores with Self-Reported Undergraduate Grades, and Related Scaled Score Means and Standard Deviations, October 1975

	Humanities and Social Science Samples N = 4447-4562			Biological and Physical Science Samples N = 2108-2214		
	Range of			Range of		
Question Type	Correlations	Means	Standard Deviations	Correlations	Means	Standard Deviations
Total Verbal	.351-.400	517-523	126-128	.259-.320	519-526	117-121
Quantitative	.233-.289	500-504	121-124	.284-.322	598-607	119-126

Table 35: Correlations of Verbal and Quantitative Ability Scores with Self-Reported Undergraduate Grades, and Related Scaled Score Means and Standard Deviations, December 1975

	Samples of Humanities, Social Science and Science Majors Combined N = 5,310-5,350		
	Range of		
Question Type	Correlations	Means	Standard Deviations
Total Verbal	.307-.338	516-520	113-114
Quantitative	.226-.266	524-528	137-138

The foregoing validity coefficients are probably attenuated by error resulting from the self-report nature of the criterion, from the inconsistency of grading practices among departments and institutions, and from lumping different samples of students together.

Population Validity

There is another type of validity that warrants special consideration. The previous discussion and the evidence presented about the validity of the GRE concern the total population of test takers. It is increasingly recognized, however, that the total population includes a variety of significant subgroups: women, men, blacks, Chicanos, foreign students, older students, and so on. It is frequently true that such subgroups have special characteristics that may render a test more or less appropriate for them. Therefore, a test may be generally valid, but may have, in one sense or another, limited validity for some prominent group of examinees. Concern for and evaluation of such possibilities can be usefully expressed as population validity.

The term *population validity* has been used in reference to the generalizability of research findings across different population groups (Bracht & Glass, 1968; Messick & Barrows, 1972). Messick (1975) pointed out that the generality of meaning of a test score across groups is an important aspect of construct validity. Writing about the Graduate Record Examinations, Willingham (1976) used the term *population validity* as a means of relating several matters concerning test bias to the larger issue of test validity. The rationale runs as follows.

Cronbach (1971) and others have emphasized that "one validates not a test but an interpretation of data arising from a specified procedure" (p. 447). That is, validity pertains not so much to the test itself as to whether the test leads to correct inferences concerning the nature of what is measured and the implications of the measurement. In that sense, the test is not valid if it means different things for different populations or leads to inferences that are systematically in error for one group or another.

A variety of questions about validity stem from the fundamental notion of whether or not a test leads to correct inferences. Many of these questions were discussed in the previous sections, and, in theory, any such questions can be applied to any particular subgroup in the population of examinees. In practice, research on population validity has focused largely on whether the selection measures are free from bias in their content and in the accuracy of the predictions they yield. In either case, the question is not whether there are any differences in the average scores earned by different groups of students. On the average, different minority groups do often earn lower scores, but this is not inconsistent with other known facts. The content of the test is intended to represent important outcomes of the mainstream educational system and the abilities necessary to do well in that system. To the extent that many members of a minority group have suffered social and educational disadvantage, they would be expected to find the test difficult, and a well-developed test should reflect the educational disadvantage a person has experienced. The critical issue in evaluating the validity of a test for such groups is rather whether the test fairly represents the developed ability or achievement it purports to measure.

Research on population validity is hampered considerably by the difficulty of obtaining relevant data. Often the necessary information (such as subgroup identification or achievement in graduate school) is not directly available to the GRE Program staff and proves difficult to obtain from students or institutions. Another serious problem is that many groups of particular interest are represented in graduate education in limited numbers, making appropriate data all the more difficult to obtain. Due to the nature of graduate education, these problems are compounded because the logical locus for many studies is the individual department, where small classes predominate and the representation of relevant subgroups is too sparse for research purposes. Some of these problems, however, can be circumvented and a great deal can be inferred about the population validity of the GRE from the considerable amount of research evidence accumulated in recent years about several similar tests used in similar circumstances. The following paragraphs summarize briefly the principle findings of such research concerning (1) predictive bias and (2) content bias.

With respect to predictive bias, there are two important questions. One is whether admission tests are as predictive of, or as highly correlated with, college performance for minorities as for majority students. The second question is whether there is any systematic tendency for the tests to underpredict or underestimate the actual performance of minority students once they are admitted. A number of studies have been directed to these two questions in undergraduate colleges and in law schools. The results are generally quite consistent.

At the undergraduate level, Stanley (1971) reviewed predictive validities for black and white students and concluded that they are quite comparable. In 1975, Cleary, Humphreys, Kendrick, and Wesman again reviewed the situation and concluded that "the predictions within black and white colleges are comparable; within integrated colleges the usual regression equations lead to comparable predictions for black and white students" (p. 31). In a review of sex bias in selection of freshman college students, Wild (1977) found a consistent trend of underprediction of women's grades when the regression equation is developed on a combined sample of men and women. She hypothesized that the differences may be due to systematic differences in meaning of the grade-point average criterion, since men and women enter different major fields in differing proportions and different fields have different grading practices.

Linn (1975) reviewed the prediction of grades in law school by the traditional measures such as undergraduate grade-point average and Law School Admission Test scores and concluded that "the traditional predictors of law school grades are usually found to be as adequate for minority persons as for majority persons, and the use of a single prediction equation usually favors the minority group member" (p. 43). Pitcher (1975) studied the prediction of grades for female law school students in 21 law schools and found that the "results of the study would in general support the use, for either men or women, of a regression system based on data for both groups combined as long as combinations of predictors [test scores and undergraduates grades] are used" (p. 1). This result supports the hypothesis of Wild (1977), since no differences were found when multiple predictors were used and for a single field of study.

These results are confirmed by an earlier review by Linn (1973) and by more recently completed studies of women, blacks, and Chicanos by Pitcher (1977) and Powers (1977). Thus, the results of research on tests generally similar to the GRE consistently indicate that academic performance of women and minority students is predicted accurately and fairly as compared to predictions of performance of males or majority students.

Available data on the GRE are consistent with that pattern. A Cooperative Validity Studies project still in progress collected validity data for 131 minority students spread through 14 graduate departments in 3 universities. Individual correlations in such small samples are quite instable—the average N was only 9—but the median validity coefficients of .33 for GRE Aptitude Test verbal ability scores and .31 for quantitative ability scores are comparable to or slightly higher than those reported for various groups of graduate students in Table 32 on page 56. Limited analyses by sex, based on data collected in the same project, show similar validities for women and men and also for foreign and nonforeign students in the quantitatively oriented fields in which foreign students tend to be found. Positive relationships between the GRE scores of foreign students and performance in graduate school have also been found in an earlier study by Harvey and Pitcher (1963).

A second general class of research on population validity has concerned the reasonableness and fairness of test content for different subgroups. Even if no evidence of predictive bias in a test is found, incorrect inferences may be drawn about the ability of students in particular subgroups because the content of the test is somehow inappropriate for those subgroups. Such research has centered on the internal characteristics of the test, particularly the question of whether certain subsets of questions tend to be inordinately difficult for some group or whether the overall pattern of difficulty is quite different for that group as compared to the total population of examinees.

With any test, it is normally assumed that the subject matter or total domain of questions may be generally unfamiliar or harder for particular groups of examinees. This in itself does not argue measurement bias, but may well represent a history of educational disadvantage with respect to the particular subject matter of the test. But if certain clusters of questions that share some particular characteristic prove overly difficult, or if the group exhibits an unusual pattern of difficulty, one could more reasonably assert that there is a bias in the choice of questions or in the process of test construction. The bias argument would then stem from an assumption that the group in question may not have had a fair opportunity to learn the particular cluster of suspect questions, or that the test questions generally may not mean the same thing for that group. In either event, the suggestion is that the domain (that is, construct) is somewhat different for the group in question and that the use of the suspect questions may be inappropriate.

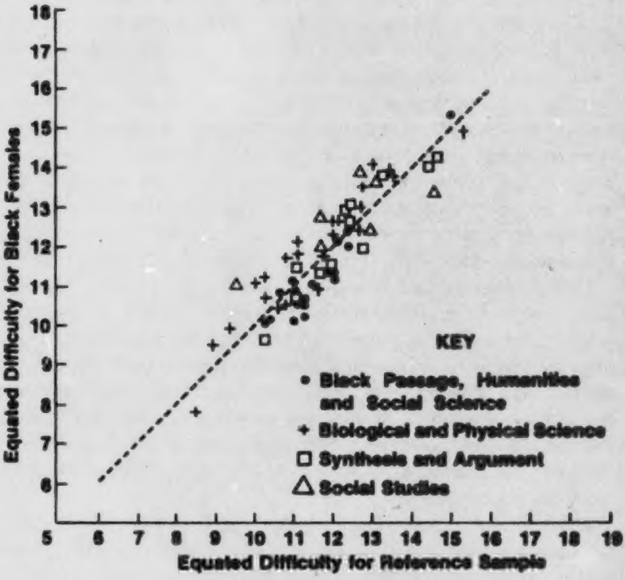
There has been a good deal of research of this general sort—often called item-group interaction studies. Again, much of the research has been done on tests generally similar to the GRE (Cleary and Hilton, 1968; Angoff and Ford, 1973; Breland et al., 1974; Swineford, 1976). If difficulties of individual questions are compared for two large groups of examinees, the difficulty of a question for one group can usually be predicted with a high degree of accuracy on the basis of its difficulty for the other group (that is, the difficulty indexes correlate on the order of .95 to .99). When such analyses have been made on the basis of samples of black and white students, a typical result is to find that practically all items in the admission test are consistently somewhat more difficult for the black group, and the correspondence of difficulty from one group to the other is still high (correlations on the order of .90), but somewhat more erratic than would be the case if two white samples were compared (Breland et al., 1974). Such a finding might suggest that different questions in the test might be perceived somewhat differently (that is, measure a different construct) for students from different cultural backgrounds, but differences in the relative difficulty of questions for different racial groups are typically not large. Furthermore, there are similar small differences between rural and urban groups and also between blacks in different cities. Finally, Angoff and Ford (1973) demonstrated that such discrepancies in relative difficulties across groups were reduced considerably by matching the two groups on overall performance; that is, regardless of ethnic group, the questions performed similarly for black and white students of high, medium, and low score levels. These findings support the assumption that the questions in these tests are perceived much the same way by black and white students and that the tests are measuring the same thing for both groups.

Another special interest in much of the research on item-group interactions has been to identify particular types of questions that stand out as significantly more difficult for women, ethnic minorities, or other groups and to avoid including such questions in the test. As yet this has not been a particularly fruitful line of inquiry. Typically, few questions stand out as unusually difficult for the subgroup, and there is usually no apparent reason for those that do. In some studies, however, it has been observed that questions associated with material having minority content tend to be somewhat easier for minority students than other questions. Due in part to such findings, a careful effort is made to include in GRE tests some material especially relevant to different subgroups, but to what extent such representation can be undertaken without changing the character of the tests or making them unfair to some other examinees is an issue not easily resolved.

Figure 7 shows an illustrative item analysis for 60 reading comprehension questions in the GRE Aptitude Test. Each point plotted in the figure represents the relative difficulty of one question for a sample of black women (N = 1,165) and for a reference sample generally representative of all examinees (N = 1,065). For illustrative purposes, the difficulty indexes (deltas) plotted here were equated so that the average delta is the same for the two groups shown. The correlation between the deltas for the two groups is quite high as correlations go (.89) but noticeably lower than correlations usually observed across comparable groups. Some consistent differences are apparent. The black women found items containing material about blacks to be relatively easier (13 items below and 4 items above the 45° equal-difficulty line), but items concerning science relatively more difficult (16 above the line and 7 below). Such results do not appear unreasonable in light of what may be assumed about learning experiences of black women, but the finding does not, in itself, warrant, or even argue for, deletion of either type of item. Both have a legitimate rationale for inclusion in the test, even though there may be small differences in the ability of one group or another to answer the different items correctly.

Information on such differences can be useful, however, in assessing the appropriateness of test content. For example, in restructuring the GRE Aptitude Test, a relatively efficient type of question—quantitative comparisons—was considered as a component of the quantitative ability measure. This type of question was shown by factor analysis to be highly related to the original types of questions in the test, and other research suggested that minority students tend to perform slightly better on this type of question than on other mathematical questions. Thus, quantitative comparisons were chosen for use because they were measures of the original construct and had content and criterion-related validity. However, this outcome tended to strengthen the case for their use because it confirmed that the change would not make the test harder for minority students.

Figure 7: Equated Difficulty of Four Types of Reading Comprehension Items in the GRE Aptitude Test for Black Females and a Representative Reference Sample



In summary, a considerable amount of research on tests like those of the GRE support the general conclusion that such examinations provide fair assessment of the particular academic abilities they represent, and that they are as predictive of the success of women and ethnic minorities as of admission applicants generally. Information available on GRE tests is consistent with that conclusion, but other research is still underway and doubtless will continue. The GRE Board's concern for the fairness of the GRE for different populations of examinees extends to other related questions that may have an important bearing on test performance—especially the possibly differential effects of coaching, speededness, and guessing habits on scores of minority groups. Research on such questions is in progress.

Validity Studies Summarized in Discussion of Predictive Validity

Alexakos, C. E. *The Graduate Record Examinations: Aptitude Tests as screening devices for students in the College of Human Resources and Education.* Unpublished report, West Virginia University, 1967. Reported by G. V. Lannholm in GRE Special Report 68-1. Princeton, N.J.: Educational Testing Service, 1968.

Besco, R. O. *The measurement and prediction of success in graduate school.* Ph.D. dissertation, Purdue University, 1960. Reported by G. V. Lannholm in GRE Special Report 68-1. Princeton, N.J.: Educational Testing Service, 1968.

Borg, W. R. GRE aptitude scores as predictors of GPA for graduate students in education. *Educational and Psychological Measurement*, 1963, *23*, 379–389.

Capps, M. P., & Decosta, F. A. Contributions of the Graduate Record Examinations and the National Teacher Examinations to the prediction of graduate school success. *Journal of Educational Research*, 1957, *50*, 383–389.

Clark, H. *Graduate Record Examination correlations with grade-point averages in the Department of Education at Northern Illinois University, 1962–1966.* Unpublished Master's thesis, Northern Illinois University, 1968.

Conway, Sister M. T. *The relationship of the Graduate Record Examination results to achievement in the Graduate School at the University of Detroit.* Unpublished Master's thesis, University of Detroit, 1955. Reported by G. V. Lannholm in GRE Special Report 68-1. Princeton, N.J.: Educational Testing Service, 1968.

Creager, J. A. *A study of graduate fellowship applicants in terms of Ph.D. attainment.* (Technical Report No. 18). Washington, D.C.: Office of Scientific Personnel, National Academy of Sciences—National Research Council, 1961.

Creager, J. A. *Predicting doctorate attainment with GRE and other variables* (Technical Report No. 25). Washington, D.C.: Office of Scientific Personnel, National Academy of Sciences—National Research Council, 1965.

Dawes, R. M. A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist*, 1971, *26*, 180–188.

Duff, F. L., & Aukes, L. E. *The relationship of the Graduate Record Examination to success in the Graduate College* (a supplementary comparative analysis of eight previously reported studies). Bureau of Institutional Research and Office of Instructional Research, University of Illinois, October 1966.

Eckhoff, C. M. Predicting graduate success at Winona State College. *Educational and Psychological Measurement*, 1966, *26*, 483–485.

Ewen, R. B. The GRE psychology test as an unobtrusive measure of motivation. *Journal of Applied Psychology*, 1969, *53*, 383–387.

Florida State University, Office of Academic Research and Planning. *The prediction of grade-point average in graduate school at the Florida State University*, Parts I & II. Florida State University, December 1971.

Florida State University, Office of Institutional Research and Service. *Relationship between Graduate Record Examinations Aptitude Test scores and academic achievement in the Graduate School at Florida State University.* Florida State University, 1958.

Hackman, J. R., Wiggins, N., & Bass, A. R. Prediction of long-term success in doctoral work in psychology. *Educational and Psychological Measurement*, 1970, *30*, 365–374.

Hansen, W. L. *Prediction of graduate performance in economics.* Department of Economics, University of Wisconsin, April 1970. (Mimeograph)

Harvey, P. R. *Predicting graduate school performance in education.* Unpublished ETS report, 1963. Reported by G. V. Lannholm in GRE Special Report 68-1. Princeton, N.J.: Educational Testing Service, 1968.

King, D. C., & Besco, R. O. The Graduate Record Examination as a selection device for graduate research fellows. *Educational and Psychological Measurement*, 1960, *20*, 853–858.

Lannholm, G. V., Marco, G. L., & Schrader, W. B. *Cooperative studies of predicting graduate school success* (GRE Special Report 68-3). Princeton, N.J.: Educational Testing Service, August 1968.

Law, A. The prediction of ratings of students in a doctoral training program. *Educational and Psychological Measurement*, 1960, *20*, 847–851.

Lorge, I. *Relationship between Graduate Record Examinations and Teachers College, Columbia University, doctoral verbal examinations* (Letter to G. V. Lannholm dated September 21, 1960). Reported by G. V. Lannholm in GRE Special Report 68-1. Princeton, N.J.: Educational Testing Service, 1968.

Madaus, G. F., & Walsh, J. J. Departmental differentials in the predictive validity of the Graduate Record Examinations Aptitude Tests. *Educational and Psychological Measurement*, 1965, *25*, 1105–1110.

Mehrabian, A. Undergraduate ability factors in relationship to graduate performance. *Educational and Psychological Measurement*, 1969, *29*, 409–419.

Michael, W. B., Jones, R. A., Al-Amir, H., Pullias, C. M., Jackson, M., & Goo, V. Correlates of a pass-fail decision for admission to candidacy in a doctoral program. *Educational and Psychological Measurement*, 1971, *31*, 965–967.

Michael, W. B., Jones, R. A., & Gibbons, B. D. The prediction of success in graduate work in chemistry from scores on the Graduate Record Examinations. *Educational and Psychological Measurement*, 1960, *20*, 859–861.

Newman, R. I. GRE scores as predictors of GPA for psychology graduate students. *Educational and Psychological Measurement*, 1968, *28*, 433–436.

Office of Educational Research. *Study of GRE scores of geology students matriculating in the years 1952–1961* (RP—Abstract, Yale University, 1963). Reported by G. V. Lannholm in GRE Special Report 68-1. Princeton, N.J.: Educational Testing Service, 1968.

Olsen, M. *The predictive effectiveness of the Aptitude Test and the Advanced Biology Test of the GRE in the Yale School of Forestry* (Statistical Report 55-6). Princeton, N.J.: Educational Testing Service, 1955. Out of print.

Roberts, P. T. *An analysis of the relationship between Graduate Record Examination scores and success in the Graduate School of Wake Forest University.* Unpublished Master's thesis, Wake Forest University, 1970.

Robertson, M., & Nielsen, W. The Graduate Record Examination and selection of graduate students. *American Psychologist*, 1961, *16*, 648–650.

Robinson, D. W. *A comparison of two batteries of tests as predictors of first year achievement in the graduate school of Bradley University*. Ph.D. dissertation, Bradley University, 1957. Reported by G. V. Lannholm in GRE Special Report 68-1. Princeton, N.J.: Educational Testing Service, 1968.

Rock, D. A. *The prediction of doctorate attainment in psychology, mathematics, and chemistry* (GRE Board Preliminary Report). Princeton, N.J.: Educational Testing Service, August 1972 (ERIC Document Reproduction Service No. ED 069 664). Later published as GRE Board Research Report 69-6aR, June 1974.

Roscoe, J. T., & Houston, S. R. The predictive validity of GRE scores for a doctoral program in education. *Educational and Psychological Measurement*, 1969, *29*, 507–509.

Sacramento State College, Test Office. *An analysis of traditional predictor variables and various criteria of success in the Master's degree program at Sacramento State College for an experimental group who received Master's degrees in the spring 1968, and a comparable control group who withdrew from their programs.* Test Office Report 69-3, Sacramento State College, October 1969.

Shaffer, J., & Rosenfeld, H. *MAT-GRE prediction study—Initial results.* Intradepartmental memorandum, Department of Psychology, University of Kansas, March 1969.

Sistrunk, F. *The GREs as predictors of graduate school success in psychology* (Letter to G. V. Lannholm dated October 3, 1961). Reported by G. V. Lannholm in GRE Special Report 68-1. Princeton, N.J.: Educational Testing Service, 1968.

Sleeper, M. L. Relationship of scores on the Graduate Record Examination to grade point averages of graduate students in occupational therapy. *Educational and Psychological Measurement*, 1961, *21*, 1039–1040.

Tully, G. E. Screening applicants for graduate study with the Aptitude Test of the Graduate Record Examinations. *College and University*, 1962, *38*, 51–60.

University of Virginia, Office of Institutional Analysis. *Correlations between admissions criteria and University of Virginia grade-point averages, Graduate School of Arts and Sciences*, Fall 1964. University of Virginia, circa 1966. (Mimeograph)

Wallace, A. D. *The predictive value of the Graduate Record Examinations at Howard University*. Unpublished Master's thesis, Howard University, 1952. Reported by G. V. Lannholm in GRE Special Report 68-1. Princeton, N.J.: Educational Testing Service, 1968.

White, E. L. *The relationship of the Graduate Record Examinations results to achievement in the Graduate School at the University of Detroit*. Unpublished Master's thesis, University of Detroit, 1954. Reported by G. V. Lannholm in GRE Special Report 68-1. Princeton, N.J.: Educational Testing Service, 1968.

White, G. W. *A predictive validity study of the Graduate Record Examinations Aptitude Test at the University of Iowa*. Unpublished Master's thesis, University of Iowa, 1967. Reported by G. V. Lannholm in GRE Special Report 68-1. Princeton, N.J.: Educational Testing Service, 1968.

Williams, J. D., Harlow, S. D., & Grab, D. A longitudinal study examining prediction of doctoral success: Grade-point average as criterion, or graduation vs. non-graduation as criterion. *Journal of Educational Research*, 1970, *64*, 161–164.

References

American Psychological Association. *Standards for educational and psychological tests.* Washington, D.C.: American Psychological Association, 1974.

Angoff, W. H., & Ford, S. F. Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, 1973, *10*, 95–105.

Bracht, G. H., & Glass, G. V. The external validity of experiments. *American Educational Research Journal*, 1968, *5*, 437–474.

Breland, H. M., Stocking, M., Pinchak, B. M., & Abrams, N. *The cross-cultural stability of mental test items* (Project Report 74-2). Princeton, N.J.: Educational Testing Service, 1974.

Brogden, H. E. On the interpretation of the correlation coefficient as a measure of predictive efficiency. *Journal of Educational Psychology*, 1946, *37*(2), 65–76.

Campbell, D. T. Recommendations for APA test standards regarding construct, trait, and discriminant validity. *American Psychologist*, 1960, *15*, 546–553.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, *56*, 81–105.

Cleary, T. A. & Hilton, T. L. An investigation of item bias. *Educational and Psychological Measurement*, 1968, *28*, 61–75.

Cleary, T. A., Humphreys, L., Kendrick, S. A., & Wesman, A. Educational use of tests with disadvantaged students. *American Psychologist*, 1975, *30*, 15–41.

Creager, J. A. *Predicting doctorate attainment with GRE and other variables* (Technical Report No. 25). Washington, D.C.: Office of Scientific Personnel, National Academy of Sciences—National Research Council, 1965.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Fishman, J. A., & Pasanella, A. K. College admission-selection studies. *Review of Educational Research*, 1960, *30*, 298–310.

Harvey, P. R., & Pitcher, B. *The relationship of Graduate Record Examinations Aptitude Test scores and graduate school performance of foreign students at four American graduate schools* (GRE Special Report 63-1). Princeton, N.J.: Educational Testing Service, April 1963.

Lannholm, G. V. *Review of studies employing GRE scores in predicting success in graduate study, 1952–1967* (GRE Special Report 68-1). Princeton, N.J.: Educational Testing Service, March 1968a.

Lannholm, G. V. *Summaries of GRE validity studies 1966–1970* (GRE Special Report 72-1). Princeton, N.J.: Educational Testing Service, February 1972.

Linn, R. L. Fair test use in selection. *Review of Educational Research*, 1973, *43*, 139–161.

Linn, R. L. Test bias and the prediction of grades in law school. *Journal of Legal Education*, 1975, *27*, 293–323.

Messick, S. The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 1975, *30*(10), 955–966.

Messick, S., & Barrows, T. S. Strategies for research and evaluation in early childhood education, in *National Society for the Study of Education, Seventy-first yearbook, Part II.* 1972, pp. 261–290.

Pitcher, B. *A further study of predicting law school grades for female law students* (Law School Admission Research Report LSAC-75-3). Princeton, N.J.: Law School Admission Council, 1975.

Pitcher, B. Subgroup validity study. Report #LSAC-76-6. In Law School Admission Council, *Reports of LSAC Sponsored Research: Volume III, 1975–1977.* Princeton, N.J.: Law School Admission Council, 1977.

Powers, D. E. Comparing predictions of law school performance for black, Chicano, and white law students. Report #LSAC-77-3. In Law School Admission Council, *Reports of LSAC Sponsored Research: Volume III, 1975–1977.* Princeton, N.J.: Law School Admission Council, 1977.

Powers, D. E., Swinton, S. S., & Carlson, A. B. *A factor analytic study of the GRE Aptitude Test* (GRE Board Professional Report 75-11P). Princeton, N.J.: Educational Testing Service, August 1977.

Rock, D. A. *The prediction of doctorate attainment in psychology, mathematics, and chemistry* (GRE Board Preliminary Report). Princeton, N.J.: Educational Testing Service, August 1972 (ERIC Document Reproduction Service No. ED 069 664). Later published as GRE Board Research Report 69-6aR, June 1974.

Stanley, J. C. Predicting college success of the educationally disadvantaged. *Science*, 1971, *171*, 640–647.

Swineford, F. Comparisons of black candidates and Chicano candidates with white candidates (LSAC Report-72-6). In Law School Admission Council, *Reports of LSAC Sponsored Research: Volume II, 1970–1974.* Princeton, N.J.: Law School Admission Council, 1976, pp. 261–263.

Wild, C. L. Statistical issues raised by Title IX requirements on admission procedures. *Journal of the National Association for Women Deans, Administrators, and Counselors*, 1977, *40*, 53–56.

Willingham, W. W. Predicting success in graduate education. *Science*, 1974, *183*, 273–278.

Willingham, W. W. *Validity and the Graduate Record Examinations.* Princeton, N.J.: Educational Testing Service, 1976.

Wilson, K. M. *GRE cooperative validity studies project: Extended progress report (75-8).* Unpublished report to the GRE Board Research Committee, 1977.

APPENDIX I

Four Types of Questions Studied but Not Selected for Use in the Analytical Ability Measure of the GRE Aptitude Test

Letter Sets

ABCDEFGHIJKLMNOPQRSTUVWXYZ

***Directions:* Each problem consists of five groups of letters. You are to find, for each group of letters, a pattern that depends *only* on the relative order in the alphabet of the letters in that group. Then choose the one group whose letters do *not* show the same pattern as that shown by the letters in the other groups.**

1. (A) ABCD (B) DEFG (C) JKLN
(D) VWXY (E) PQRS

The pattern in groups (A), (B), (D), and (E) can be summarized as follows: The letters in the group are in consecutive alphabetical order. The letters in the group (C) do *not* show this pattern. Therefore, the correct answer is (C).

Each group of letters should be considered independently. Look at relationships within groups instead of relationships between groups. Do not concern yourself with whether the letters fall near or at the beginning or end of the alphabet, or with whether the letters are vowels or consonants. Do not consider differences in the sounds represented by the letters, or the relationship of the letters to any other groups of letters, such as words. Consider *only* the relative order in the alphabet of the letters in a group.

2. (A) CEGI (B) EGIK (C) GIKM
(D) IKMP (E) PRTV

The pattern in groups (A), (B), (C), and (E) can be summarized as follows: Each of the four letters is separated from the next by one successive letter. The letters in group (D) do not show this pattern, so (D) is the correct answer. (Note: It would be wrong to choose answer (E) on the basis of the statement: All of the other groups contain the letter I. This statement has nothing to do with the relative order of letters within each of several independent groups.)

Logical Reasoning

***Directions:* The questions in this section require you to evaluate the reasoning contained in brief statements or passages. In some questions, each of the choices is a conceivable solution to the particular problem posed. However, you are to select the one that is best, that is, the one that does not require you to make what are by common-sense standards implausible, superfluous, or incompatible assumptions. After you have chosen the best answer, blacken the corresponding space on the answer sheet.**

1. Since all rabbits that I have seen have short tails, all rabbits probably have short tails.

Which of the following most closely parallels the kind of reasoning used in the sentence above?

(A) Since all chemical reactions that I have seen have been undramatic, probably only minor changes took place in the substances involved.
(B) Since all the human social systems that I have heard of have sexual taboos, all of these sexual taboos have probably had survival value for the human race.
(C) Since all of the plays of Jovita Maldonado that I have seen feature a spurned lover, probably all of her plays feature this character type.
(D) Since all eating utensils that I have seen are made of metal, metal is probably the most desirable material for eating utensils.
(E) Since sight is the most important of man's five major senses, its failure probably seriously affects an individual's aptitude for all formal education.

The statement on which this quite easy question is based reflects inductive reasoning: generalizing about an entire class on the basis of specific observations. Although one could criticize the conclusion by pointing to the limitations of the observations, this question does not ask for an evaluation of the reasoning process but for recognition of a parallel example of that kind of reasoning.

All of the answer choices are similar in some ways, but only one is a statement about specific observations followed by a generalization based on those observations. In (A), (B), and (D), the second part of the statement is not a *generalization* based on the observations mentioned in the first part but is an *explanation* or suggested reason for what was observed. In (E) an assumption is followed by a conclusion. Only (C) refers to specific observations (about some of Maldonado's plays) and proceeds to generalize (about all of Maldonado's plays) on the basis of these observations, however limited they may be. Therefore (C) is the correct answer.

2. A good hotel can give you a beautiful room for $30 a day, with three meals, and make a profit and pay taxes. And yet a tax-exempt hospital operates in the red for $65 a day. I say it must be bad administration.

The author's argument would be considerably weakened if attention were drawn to the fact that

(A) hotel managers receive better training than do hospital administrators
(B) the quality of food served by hotels exceeds that of food served in hospitals
(C) hospitals are run by dishonest administrators
(D) hospitals provide other services besides room and board
(E) hospital deficits are a recent phenomenon

This very easy question focuses on the reasonableness of drawing a conclusion from the evidence presented. The author's contention is based on some evidence—the discrepancy between a hotel's and a hospital's operating expenses. The question asks you to identify additional evidence that would weaken the argument that bad administration is responsible for the discrepancy in expenses. (A) and (B) cannot be that evidence because these choices, if true, would actually strengthen the author's argument. (C) is a slightly altered version of the author's own statement of a reason for the discrepancy. (E) could weaken the argument, but only if more information were given. Only (D) provides evidence that casts doubt on the argument. If hospitals provide services other than those mentioned, then the costs of those services rather than bad administration are likely to be the reason for the difference between hotel and hospital expenses. Therefore (D) is the correct answer.

***Questions 3–4* refer to the following passage.**

A servant who was roasting a stork for his master was prevailed upon by his sweetheart to cut off one of its legs for her to eat. When the bird was brought to the table, the master asked what had become of the other leg. The man answered that storks never had more than one leg. The master, very angry but determined to render his servant speechless before he punished him, took the servant the next day to the fields where they saw storks, each standing on one leg. The servant turned triumphantly to the master; but the master shouted, and the birds put down their other legs and flew away. "Ah, sir," said the servant, "you did not shout to the stork at dinner yesterday; if you had, he too would have shown his other leg."

3. The servant's final retort to his master would be true if which two of the following statements were simultaneously true?

I. Roasted storks at the dinner table behave just as live storks in the field do.
II. The missing leg on yesterday's roasted stork had actually been tucked under the bird.
III. The master had not undertaken to teach the servant a lesson.
IV. The servant's sweetheart, rather than the servant himself, had cut off the stork's leg.

(A) I and II (B) I and III (C) II and III
(D) II and IV (E) III and IV

The humor of the fable on which these questions are based derives from a logical problem relating to the drawing of conclusions from evidence. The first question is based on the servant's clever retort to his master. The servant's retort is a logical conclusion if certain assumptions are made. The servant's argument—that if the master had shouted at the roasted stork on yesterday's table, it would have shown its other leg as did the storks in the field—assumes that roasted and live storks behave in the same way (I) and also assumes what the servant would like the master to believe, that the missing leg on yesterday's roast had been tucked under the bird (II) rather than eaten by the servant's sweetheart. If (III) were true, the situation leading to the servant's retort would not have occurred; but (III) does not bear on the truth of the servant's retort. (IV) is not the answer because it does not matter who cut off the stork's leg: what is important to the situation is that it was cut off. Thus (A), (I and II), is the correct answer. This question is of about average difficulty.

4. The servant was able to attack the master's demonstration primarily because the master failed to

(A) objectively consider the possibility that storks might have only one leg each
(B) take anyone with him to the fields to confirm his observations
(C) plan later experiments to follow up and validate his tentative findings
(D) reveal to his servant that it was possible for storks to lack one leg and still fly
(E) consider that conditions governing the demonstration were unlike those of the previous day's happening

This question, a moderately easy one, asks for a criticism of the master's demonstration. The master did prove what he had intended to prove—that storks have more than one leg. However, the servant was able to elude the master's "proof" of his guilt by immediately accepting the two-leggedness of storks (including that of a roasted stork) but claiming that the master had simply not treated the roasted and live storks in the same way. Thus (E) is correct because the master's experiment took place under different conditions than did the previous day's experience. (D) is irrelevant to the argument and can be easily eliminated. Choices (A), (B), and (C) all sound as if they might be flaws in an experiment intended to be scientific. But they do not explain why the servant was able to spring back with a new argument. Objective consideration of the possibility that storks might have only one leg each, (A), would not have strengthened the master's demonstration. More observers, (B), or additional experiments, (C), to confirm his findings would perhaps have strengthened his point about live storks but would have had no implications for roasted ones. Thus only (E) can be the answer.

***Questions 5–6* refer to missing portions of the following passage. For each question, choose the completion that is best according to the context of the passage.**

If a book disgusted everyone, no one would read it. However, one can be sure of selling many copies of a book that is publicly proclaimed obscene, for the officially held standards of propriety do not prevail throughout the community. At this point I may be expected to denounce the hypocrisy of the age. I shall not do so. The concept of hypocrisy applies to morals: a person should be good and not merely seem so, and a bad person is little mended by pretense of goodness. But propriety is altogether a matter of how actions appear, so that 5. If a person seems to give no offense, he gives no offense. Why, then, should a society not have public standards of propriety different from those applied by each citizen to his own private conduct? It would be no more absurd to advertise filthy movies by decorous posters than it is to advertise decorous movies by filthy posters; and if a society in which everyone avidly read pornography were to forbid its public sale, that would mean only that it combined a taste for such reading with a taste for 6.

5. (A) bad men rarely succeed in appearing good
(B) the concept of hypocrisy does not apply
(C) actions appear different to observers with different standards
(D) the issue is basically a moral one
(E) the concepts of impropriety and immorality are indistinguishable

Both this question and the next require that you follow the author's reasoning well enough to fill in missing material. The first question focuses on the distinction that the author makes between morality and propriety. The author suggests that propriety, unlike morality, is entirely a matter of appearance and has nothing to do with what is really good or bad. *Because the author contrasts morality* and propriety, (D) and (E) can be eliminated. (D) assumes that propriety is a moral issue. (E) suggests that impropriety and immorality (and by inference propriety and morality) are indistinguishable. Since "bad" and "good" in (A) refer to morality, (A) does not follow from the author's ideas on propriety. (C) is an appealing answer, since it focuses on the way actions appear in public. But (C) does not follow from the words "so that," leading from the first clause ("propriety is altogether a matter of how actions appear") to the second clause. Only (B) is an acceptable answer. It follows from the author's distinction between propriety and morality because it states that hypocrisy (associated with the concept of morality) does not apply to the question of propriety. This question is difficult.

6. (A) obscenity (B) hypocrisy (D) oppression
(D) good art (E) decorum

The answer to this very difficult question must meet two requirements. It must describe something that is consistent, in the context of the author's discussion, with the reading of pornography. It must also explain the paradox, described in the preceding lines, of a society in which pornography is read although public sale of pornography is forbidden. The sense of the author's argument is that it is possible not to call such a society hypocritical (lines 4-9). Therefore (B) is a poor choice. (A), "obscenity," is a term similar to "pornography" and does not suggest any contrast that could explain the paradox. "Oppression" and "good art," though relevant in general to the topic of pornography, are not relevant to the author's discussion here. Thus (C) and (D) can be eliminated. (E), "decorum," a synonym for "propriety," does fit into the context. It explains the contrast between privately reading pornography and publicly banning it, according to the author's view that propriety has to do with public appearance only and not with private actions. Thus (E) is the correct answer.

Evaluation of Evidence

Directions: **Each of the sets in this section consists of a description of a fact situation and a conclusion based on that situation. Following each conclusion are a number of statements.**

Consider each statement separately **in relation to the fact situation. Then on the answer sheet blacken space**

A if the statement either proves the conclusion or makes it almost certainly true;
B if the statement supports the conclusion but does not make it almost certainly true;
C if the statement either disproves the conclusion or makes it almost certainly false;
D if the statement weakens the conclusion but does not make it almost certainly false;
E if the statement is irrelevant to the conclusion or affects it only slightly.

Sample Set

Between 10:00 p.m. and 11:00 p.m. on October 31, five children were admitted to the Fairchild General Hospital. Four were suffering from severe stomach cramps and vomiting. All five had been out "trick or treating" and had eaten a good deal of candy. Doctors at the hospital diagnosed cyanide poisoning and called the police. The police ascertained that only one street, Mavis Avenue, had been canvassed by all five of the children, three in one group and two in another. Their bags of candy were impounded.

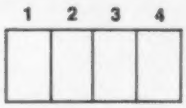
When the residents of Mavis Avenue were interviewed, several mentioned that John Ames, their neighbor, had said, "I'm going to give some kids a Halloween they won't forget." Records at the corner pharmacy indicated that Ames had purchased cyanide on October 29.

Conclusion: **Ames poisoned the children.**

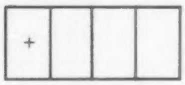
		Sample Answers
Sample 1:	**Some of the candy remaining in the children's bags contained cyanide. Ames's fingerprints were found on the wrappers of the poisoned candy.**	● Ⓑ Ⓒ Ⓓ Ⓔ
Sample 2:	**All five children remembered going to the Ames house.**	Ⓐ ● Ⓒ Ⓓ Ⓔ
Sample 3:	**Ames had given pennies, not candy, to the children.**	Ⓐ Ⓑ ● Ⓓ Ⓔ
Sample 4:	**Three of the five hospitalized children did not think they had gone to the Ames house.**	Ⓐ Ⓑ Ⓒ ● Ⓔ
Sample 5:	**Several of Ames's friends said that they would vouch for his character.**	Ⓐ Ⓑ Ⓒ Ⓓ ●

Deductive Reasoning

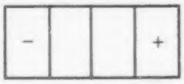
Directions: **The questions in this section are based on diagrams consisting of a rectangle divided into 4 regions.**



A plus sign (+) in a region represents the statement that there is something in the region. A minus sign (−) in a region represents the statement that there is nothing in the region.

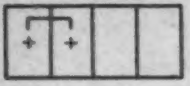


There is something in the first region.

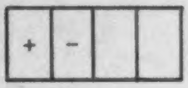


There is nothing in the first region, and there is something in the fourth region.

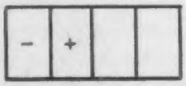
If a bracket (┌┐) connects two signs, then one of the signs holds and the opposite of the other sign holds.



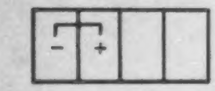
is equivalent to either



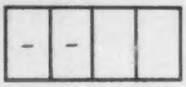
or



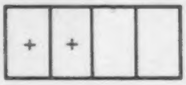
If two plus signs or two minus signs are connected by a bracket, then one plus sign and one minus sign must result. If a plus sign and a minus sign are connected by a bracket, then two plus signs or two minus signs must result.



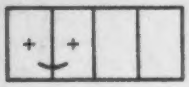
is equivalent to either



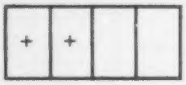
or



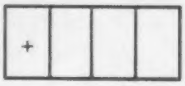
If an arc (‿) connects two signs, then AT LEAST ONE of the signs holds, and BOTH signs may hold.



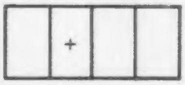
is equivalent to either



or

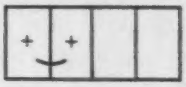


or

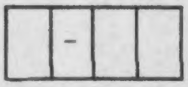


If two diagrams are given, the information in them may be combined in a single diagram.

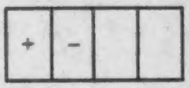
Given



and



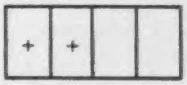
then



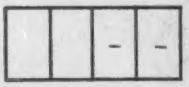
must result.

Sample Question

Given

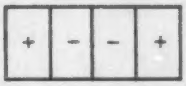


and

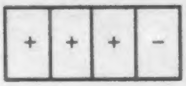


Which of the following can result?

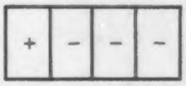
(A)



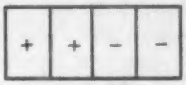
(B)



(C)



(D)



D is the correct answer.

APPENDIX II

Information Unique to Each Advanced Test

On the following pages, the content of each of the 20 Advanced Tests is described. In addition, for all the tests except Advanced Computer Science, the percentage distributions of students' responses to questions about their field backgrounds and educational goals are reported. These data are based on the responses of students taking the Advanced Tests at all administrations during the 1970–71 academic year. The percentage not responding at all to each question is not given.

Validity data are also summarized where available. Only studies using an Advanced Test as one of the predictors and involving students at least some of whom entered graduate school as late as 1956 are included. Three persistent problems make it difficult to interpret validity coefficients: 1) small samples of students, 2) inadequate criteria of success in graduate school, and 3) restriction of the range of measures for both predictors and criteria. (A full discussion of validity is presented in Chapter 6.) For each validity study, the student group involved is described in terms of size, institution, year, and other pertinent characteristics, as available. The predictors and criteria and the relations between these variables are reported.

ADVANCED BIOLOGY TEST

Content

To cover the broad field of the biological sciences, the subject matter is organized into the three major areas of cellular and subcellular, organismal, and population biology.

About one-third of the examination is concentrated on the materials and phenomena found at the subcellular and cellular levels of organization. "Subcellular" is defined operationally to include atomic and molecular species, macromolecules, and such structures as cell organelles and viruses. "Cellular," also defined operationally, includes unicellular organisms and the distinctive cell types of multicellular organisms. Under this general heading, consideration is given to the chemistry and physics of the atoms and molecules found in biological systems as well as to their functions and architectural involvement. The energetics of subcellular and cellular levels is included, emphasizing photosynthesis, synthetic and degradative pathways, and maintenance needs. Homeostatic mechanisms are examined from the metabolic and stimulus-response aspects. Replicative processes and the means of transmission of information for them are considered, and some attention is given to techniques of study.

The organismal biology questions are concerned with the biology of multicellular organisms as individuals. The questions relate to the genetic and environmental control of growth and development, structure and function, and behavior. Development includes those processes from fertilization through organogenesis to postembryonic development, metamorphosis, senescence, regeneration, life cycles, and transmission of hereditary characters. Structural and functional aspects include homeostatic mechanisms at the tissue and organ levels and hormonal and neural integration. Behavior encompasses reflex mechanisms, spontaneous activity, innate and motivated behavior, biorhythms, maturational changes, and various forms of learning.

The population biology questions deal with populations and their responses to environmental factors and genetic change. Included are the systematics of organisms and ecosystem structure and function, with consideration of such topics as energy flow, material cycling, community homeostasis, and the ecological impact of human activity. Other aspects of population biology considered are population genetics, population behavior, the evolutionary sequence of organisms, and the mechanisms by which evolution has occurred.

Since certain abilities are judged important in undergraduate biology curriculums, consideration is given to evaluating the student's

- understanding of (a) the historical development of basic biological concepts and (b) scientific processes and methods of investigation, including recognition of the tentative character of much scientific knowledge;
- ability to apply the techniques and methods of biological science to the interpretation of laboratory and field situations and basic research findings;
- ability to use resource material, evaluate unfamiliar material, and establish relationships between the contributions of biological science and those of other disciplines.

Responses to Background Questions, 1970–71 (N = 13,496)

A. At what point are you in your studies?

3% (1) I am in or have just completed my junior year of undergraduate study.
67% (2) I am an undergraduate senior.
20% (3) I have a bachelor's degree but am not presently enrolled in graduate school.
8% (4) I am in or have just completed my first year of graduate study.
7% (5) I am in or have completed my second year of graduate study.

B. What graduate degree do you intend to seek?

4% (1) I do not plan to pursue graduate study.
3% (2) I plan to pursue graduate work but not to obtain a graduate degree.
33% (3) I plan to obtain terminal M.A., M.S., or other degree at the master's level.
31% (4) I plan to obtain M.A., M.S., or other master's level degree leading to a doctoral degree.
28% (5) I plan to obtain Ph.D., Ed.D., or other degree at the doctoral level.

C. How many semester (quarter) courses of chemistry did you take as an undergraduate?

1% (1) None
11% (2) Two or fewer
44% (3) Three or four
30% (4) Five or six
14% (5) Seven or more

D. How many semester (quarter) courses of physics did you take as an undergraduate?

18% (1) None
16% (2) One
49% (3) Two
12% (4) Three
5% (5) Four or more

E. How many semester (quarter) courses of mathematics did you take as an undergraduate?

5% (1) None
15% (2) One
40% (3) Two
20% (4) Three
19% (5) Four or more

F. If you were an undergraduate biology major, in which of the following areas did you specialize?

39% (1) General Biology
25% (2) Zoology
5% (3) Botany
6% (4) Microbiology
18% (5) Other

Validity Data

1. The subjects for a study by Creager (1965) were 460 applicants (320 males and 140 females) for National Science Foundation fellowships in 1955 and 1956. The predictors were scores on the GRE Aptitude Test (verbal and quantitative) and the Advanced Biology Test. One criterion was time lapse between attainment of the B.A. and the Ph.D., coded as shown below:

B.A.-Ph.D. Time Lapse (in years):	Less than 4	4	5	6	7	8	9	No Ph.D. by Aug. '64
Coded Variable:	1	2	3	4	5	6	7	8

A second criterion was the dichotomous variable of attaining or not attaining a Ph.D. by August 1964. The third criterion was the dichotomous variable of attaining or not attaining a Ph.D. in the average time taken to attain a Ph.D. in the field. The relationships between predictors and criteria are shown in Tables 1 and 2.

Table 1: Validities of GRE Scores against Doctorate Attainment for 320 Males Who Were Applicants for National Science Foundation Fellowships in Biology in 1955 and 1956

	Criteria				
		Ph.D. by 1964		Ph.D. in Average Time	
Predictors	Reflected B.A.-Ph.D. Time Lapse[1]	Point Biserial	Biserial	Point Biserial	Biserial
GRE Verbal Ability	.23	.20	.26	.20	.26
GRE Quantitative Ability	.23	.21	.27	.21	.27
GRE Advanced Biology	.18	.14	.18	.17	.22
Composite	.26	.23	.29	.23	.30

Table 2: Validities of GRE Scores against Doctorate Attainment for 140 Females Who Were Applicants for National Science Foundation Fellowships in Biology in 1955 and 1956

	Criteria				
		Ph.D. by 1964		Ph.D. in Average Time	
Predictors	Reflected B.A.-Ph.D. Time Lapse[1]	Point Biserial	Biserial	Point Biserial	Biserial
GRE Verbal Ability	.14	.08	.09	.14	.22
GRE Quantitative Ability	.20	.11	.17	.22	.35
GRE Advanced Biology	.23	.17	.26	.23	.37
Composite	.25	.19	.29	.26	.41

[1] Correlations between the coded variable for B.A.-Ph.D. time lapse given above and the predictors with the signs reversed

2. Roberts (1970) studied the records of 41 students who had enrolled at Wake Forest University from June 1964 to June 1970 for graduate study in biology and who had completed at least nine hours of graduate work. The correlations between graduate grade-point averages and GRE scores were .24 for verbal ability, .27 for quantitative ability, and .36 for Advanced Biology.

ADVANCED CHEMISTRY TEST

Content

The content of the examination emphasizes the four fields into which chemistry has been traditionally divided and their interrelationships. An outline of the material covered by the test follows:

I. ANALYTICAL CHEMISTRY — 15 percent

A. Classical quantitative area

Titrimetry; separations, including theory and applications of chromatography as well as gravimetry; data handling, including statistical tests (t, F, Q, chi-square); standards and standardization techniques

B. Instrumentation area

Basic electronics; electrochemical methods; spectroscopic methods, including mass spectroscopy and those in the electromagnetic spectrum from high-energy nuclear processes of radioactivity to nuclear magnetic resonance

II. INORGANIC CHEMISTRY — 25 percent

A. Atomic theory

Elementary particles, atomic structure, classical experiments

B. The nucleus

Binding energy, abundance and stability of nuclei, isotopes

C. Extranuclear structures and related properties

Electronic distributions in atoms, periodic classifications, properties dependent on extranuclear structure

D. Chemistry of the families of elements

Preparations, reactions, properties, and important applications of the elements and their compounds stressing family relationships and dependence on extranuclear structure; families of representative elements, families of transition elements, lanthanides and actinides

III. ORGANIC CHEMISTRY — 30 percent

A. Principal reactions of simple functional groups

Hydrocarbons, alcohols, alkyl and aryl halides, organometallic compounds, carbonyl compounds, conjugate unsaturated carbonyl compounds, amines, diazonium compounds, acids, phenols, simple sulfur-containing compounds

B. Structure and mechanism

Electronic structures, isomers and stereochemistry, theoretical concepts, basic reaction mechanisms, structural interpretation of spectral (ultraviolet, infrared, nuclear magnetic resonance) data

C. More advanced topics and special topics

Laboratory topics, classical reaction types, classical rearrangements, differentiations by chemical tests, special reagents, bifunctional compounds, polymerizations, natural products, comparisons of reactivity, biochemically related topics

IV. PHYSICAL CHEMISTRY — 30 percent

A. Classical and statistical thermodynamics

Equations of state; first, second, and third laws; E(U), H, S, G, μ, C_p, C_v; phase equilibria; equilibrium conditions; Nernst's equation; elementary statistical mechanics

B. Quantum chemistry and spectroscopy

Energy levels and wave functions for atomic and molecular electrons, harmonic oscillators, rigid rotors, and translational motion; selection rules; microwave, infrared, visible, Raman, and nuclear magnetic resonance spectroscopy

C. Kinetics and other topics

Elementary kinetic theory of gases; rate laws and mechanisms; crystallography; dielectric properties; electrochemistry; surface chemistry; polymers; chemistry of solutions; applications to biological systems

Each form of the examination samples widely among the topics listed above, but questions on all the topics are not in every examination.

Responses to Background Questions, 1970–71 (N = 5,126)

A. At what point are you in your studies?

3% (1) I am in or have just completed my junior year of undergraduate study.
65% (2) I am an undergraduate senior.
17% (3) I have a bachelor's degree but am not presently enrolled in graduated school.
7% (4) I am in or have just completed my first year of graduate study.
7% (5) I am in or have completed my second year of graduate study.

B. What graduate degree do you intend to seek?

3% (1) I do not plan to pursue graduate study.
2% (2) I plan to pursue graduate work but not to obtain a graduate degree.
19% (3) I plan to obtain terminal M.A., M.S., or other degree at the master's level.
24% (4) I plan to obtain M.A., M.S., or other master's level degree leading to a doctoral degree.
50% (5) I plan to obtain Ph.D., Ed.D., or other degree at the doctoral level.

Validity Data

1. A group studied by Rock (1972) included 643 applicants for National Science Foundation fellowships. Most applied for first-year NSF fellowships in 1958–61. The predictors were scores on the GRE Advanced Chemistry Test and GRE Aptitude Test (verbal and quantitative), undergraduate grade-point average, and an average rating of reference letters. The criterion was attainment of the doctorate by June 1968. The group was split into random halves; the validity coefficients for each half are shown in Table 3.

Table 3: Validities Using the Criterion of Attainment of the Doctorate for 643 National Science Foundation Fellowship Applicants in Chemistry in 1958–61, Split into Two Random Halves

	n = 322			n = 321		
Predictors	r-biserial Correlation with Criterion	Predictor Performance		r-biserial Correlation with Criterion	Predictor Performance	
		Mean	Standard Deviation		Mean	Standard Deviation
GRE Advanced Chemistry[1]	0.33	67.41	11.81	0.48	66.27	12.31
GRE Verbal Ability[1]	0.15	59.66	10.69	0.23	58.40	10.75
GRE Quantitative Ability[1]	0.28	69.26	10.70	0.34	67.96	10.70
Undergraduate GPA[2]	0.27	246.93	44.55	0.36	247.93	42.82
Reference Letters[3]	0.30	41.48	9.89	0.33	42.23	9.15

[1] Scaled score with the third digit dropped
[2] On a four-point scale multiplied by 100
[3] Zero to 6 multiplied by 100

2. The subjects for a study by Creager (1965) were 660 applicants (500 males and 160 females) for National Science Foundation fellowships in 1955 and 1956. The predictors were scores on the GRE Aptitude Test (verbal and quantitative) and the Advanced Chemistry Test. One criterion was time lapse between attainment of the B.A. and Ph.D., coded as shown below:

B.A.-Ph.D. Time Lapse (in years):	Less than 4	4	5	6	7	8	9	No Ph.D. by Aug. '64
Coded Variable:	1	2	3	4	5	6	7	8

A second criterion was the dichotomous variable of attaining or not attaining a Ph.D. by August 1964. The third criterion was the dichotomous variable of attaining or not attaining a Ph.D. in the average time taken to attain a Ph.D. in the field. The relationships between predictors and criteria are shown in Tables 4 and 5

Table 4: Validities of GRE against Doctorate Attainment for 500 Males Who Were Applicants for National Science Foundation Fellowships in Chemistry in 1955 and 1956

	Criteria				
		Ph.D. by 1964		Ph.D. in Average Time	
Predictors	Reflected B.A.-Ph.D. Time Lapse[1]	Point Biserial	Biserial	Point Biserial	Biserial
GRE Verbal Ability	.16	.12	.15	.13	.16
GRE Quantitative Ability	.26	.21	.26	.22	.28
GRE Advanced Chemistry	.38	.31	.39	.34	.43
Composite	.39	.33	.41	.35	.44

[1] Correlations between the coded variable for B.A.-Ph.D. time lapse given above and the predictors with the signs reversed

Table 5: Validities of GRE against Doctorate Attainment for 160 Females who Were Applicants for National Science Foundation Fellowships in Chemistry in 1955 and 1956

	Criteria				
		Ph.D. by 1964		Ph.D. in Average Time	
Predictors	Reflected B.A.-Ph.D. Time Lapse[1]	Point -Biserial	Biserial	Point Biserial	Biserial
GRE Verbal Ability	.15	.17	.25	.17	.25
GRE Quantitative Ability	.29	.27	.39	.27	.39
GRE Advanced Chemistry	.38	.37	.37	.37	.54
Composite	.40	.38	.38	.38	.55

[1] Correlations between the coded variable for B.A.-Ph.D. time lapse given above and the predictors with the signs reversed

3. Roberts (1970) studied the records of 31 students who had enrolled at Wake Forest University from June 1964 to June 1970 for graduate study in chemistry and who had completed at least nine hours of graduate work. The correlations between graduate grade-point averages and GRE scores were .55 for verbal ability, –.07 for quantitative ability, and .11 for Advanced Chemistry.

ADVANCED COMPUTER SCIENCE TEST

Content

The approximate distribution of questions in each edition of the test according to content categories is indicated by the following outline. The items in parentheses are intended to be examples of topics under the headings, not exhaustive lists. The percentages given are approximate; actual percentages will vary slightly from one edition of the test to another. The issue of how to balance theory versus practical applications is a problematical one and is not yet completely resolved.

I. Programming Systems and Methodology — 40 percent

- A. Programming Languages and their Processors

  (Evaluation of expressions, block structure, parameter passing and binding, control structures, assemblers, compilers, interpreters)

- B. Programming Concepts

  (Iteration, recursion, modularity, abstraction, refinement, verification, documentation)

- C. Properties of Algorithms

  (Time and space requirements of programs, especially of common processes such as sorting and searching; correctness of programs)

- D. Data Structures

  (Linear data structures, list structures, strings, stacks, queues, trees)

- E. Operating Systems

  (Scheduling, resource and storage allocation, interrupts, synchronization, addressing techniques, file structures, editors, batch/time sharing, networks/communications)

II. Computer Systems — 20 percent

- A. Logic Design

  (Switching algebra, combinatorial and sequential networks)

- B. Implementation of Computer Arithmetic

  (Codes, number representation, add/subtract/multiply/divide)

- C. Processor Organization

  (Instruction sets, registers, data and control flow, storage)

- D. System Architecture

  (Configurations of and communication among processors, memories, and I/O devices)

III. Theory of Computation — 15 percent

- A. Automata Theory

  (Sequential machines, transitions, regular expressions, Turing machines, nondeterministic finite automata)

- B. Analysis of Algorithms

  (Complexity of specific algorithms, exact/asymptotic/lower bound analysis, analysis of time/space complexity, correctness)

- C. Formal Languages

  (Regular and context-free grammars/languages, simple properties such as emptiness or ambiguity)

IV. Computational Mathematics — 20 percent

- A. Discrete Structures

  (Logic, sets, relations, functions, Boolean algebra, linear algebra, graph theory, combinatorics)

- B. Numerical Mathematics

  (Arithmetic, number representation, numerical algorithms, error analysis, discrete probability, and elementary statistics)

V. Special Topics — 5 percent

(Simulation and modeling, data management systems, information retrieval, artificial intelligence)

Since the Advanced Computer Science Test was only recently introduced, answers to background questions by students and validity data are not available.

ADVANCED ECONOMICS TEST

Content

The committee responsible for the test has felt that the primary concern in graduate school selection is the student's competence in the basic skills of economic analysis rather than his or her knowledge of economic institutions. Thus, the test's content specifications have increasingly emphasized basic macro- and microeconomic analysis.

In current editions approximately 60 percent of the test is evenly divided between macroeconomic and microeconomic analysis. The remainder consists of questions on broad topics that might be covered in a variety of upperclass economics courses, including money and banking, international economics, labor, industrial organization, public finance, quantitative economics, comparative economic systems, urban economics, and economic development. Although an individual question may be couched, for example, in terms of international trade, the information needed to answer it often has been studied in several other courses.

An important consideration in planning the content of the test is the emphasis to be given to subjects covered in such courses as those described above—courses other than introductory principles or macro- and microeconomics. Although a good proportion of undergraduate economics majors study money and banking, international trade, and public finance, it is recognized that substantial emphasis upon any such topics would penalize those who have not taken courses devoted to them. Since the preparation of students varies, it is expected that each score will be evaluated in light of the student's record as of the time he or she takes the test.

Although questions in quantitative economics are generally difficult for many students currently taking the test, some such questions are included because of the increasing importance of the subject and its relevance to success in many graduate schools.

Responses to Background Questions, 1970–71 (N = 4,770)

A. At what point are you in your studies?

3% (1) I am in or have just completed my junior year of undergraduate study.
61% (2) I am an undergraduate senior.
18% (3) I have a bachelor's degree but am not presently enrolled in graduate school.
9% (4) I am in or have just completed my first year of graduate study.
6% (5) I am in or have completed my second year of graduate study.

B. What graduate degree do you intend to seek?

6% (1) I do not plan to pursue graduate study.
2% (2) I plan to pursue graduate work but not to obtain a graduate degree.
36% (3) I plan to obtain terminal M.A., M.S., or other degree at the master's level.
24% (4) I plan to obtain M.A., M.S., or other master's level degree leading to a doctoral degree.
30% (5) I plan to obtain Ph.D., Ed.D., or other degree at the doctoral level.

C. If you are now a college senior, which of the following best describes your educational experience and your plans with respect to the graduate study of economics? (If you are not a senior, mark 5.)

34% (1) I am an undergraduate major in economics and I plan to do graduate work in economics.
17% (2) I am an undergraduate major in economics and I plan to do graduate work in an area related to economics.
49% (3) I am an undergraduate major in economics, but I plan to do graduate work in an area unrelated to economics.
7% (4) I am not an undergraduate major in economics, but I plan to do graduate work in economics.
30% (5) Not a senior, or other

D. If you were not an undergraduate major in economics, in which of the following areas did you specialize?

30% (1) Social Sciences (including Business)
4% (2) Engineering
2% (3) Biological or Physical Sciences
7% (4) Mathematics (including Statistics)
4% (5) Humanities

In questions E through I, include in your answer the courses in which you are currently enrolled only if you have completed more than half a term.

E. How many semester (quarter) courses beyond the first course have you had in microeconomic analysis (the study of individual economic units—markets, firms, consumers, workers)?

24% (1) None
40% (2) One
18% (3) Two
8% (4) Three
6% (5) Four or more

F. How many semester (quarter) courses have you had beyond the first course in macroeconomic analysis (the study of aggregate economic behavior including monetary theory)?

23% (1) None
37% (2) One
21% (3) Two
8% (4) Three
8% (5) Four or more

G. How many semester (quarter) courses have you taken in economic statistics and econometrics?

37% (1) None
36% (2) One
16% (3) Two
5% (4) Three
3% (5) Four or more

H. How many other semester (quarter) courses in economics have you had including the introductory course(s)?

2% (1) None
10% (2) One or two
18% (3) Three or four
24% (4) Five or six
42% (5) Seven or more

I. How many semester (quarter) courses in mathematics (including mathematical probability and statistics) have you taken?

35% (1) Two or fewer
34% (2) Three or four
9% (3) Five
9% (4) Six or seven
9% (5) Eight or more

ADVANCED EDUCATION TEST

Content

Questions are drawn from the courses of study most commonly offered. Since the emphasis is placed on the relationships among the content dimensions of education, the particular pattern of courses students have taken is likely to be less crucial than their ability to integrate the knowledge and skills they have gained.

The test questions, for the most part, ask the student to solve problems using basic concepts, knowledge, understanding, and abilities from those areas from which the substantive content of education is generally drawn—that is, history, philosophy, psychology, and sociology. Various concerns in education are considered. These concerns and the relative weight of each in the test are (1) educational goals, 15 percent; (2) administration and supervision of the schools, 15 percent; (3) curriculum development and organization, 15 percent; (4) teaching-learning, 40 percent; (5) evaluation and research appraisal, 15 percent. The following outline provides greater detail.

1. Educational goals, including (a) the aims of education and their proponents and justification, viewed philosophically and in historical perspective; (b) the clarification and feasibility of a variety of possible goals of education, with particular reference to physical, emotional-social, and intellectual development; (c) the role of education as related to a pluralistic society, community goals, social problems, and so on.
2. Administration and supervision of the schools, including (a) sources of influence and authority, viewed historically and philosophically; (b) psychological considerations, such as grouping for learning and staff and student morale; (c) the teacher's legal rights and responsibilities; sociological considerations, such as community characteristics, needs, aspirations, and role in educational planning.
3. Curriculum development and organization, including (a) evolution of the cirriculum in the schools and philosophical dimensions of curriculum issues; (b) curriculum as related to stages of growth and development and learning factors; (c) curriculum as related to societal demands on education, and so on.
4. Teaching-learning, including (a) the evolution of theories of teaching-learning and their relationship to curriculum types and teaching styles; logical aspects of teaching, including defining, explaining, questioning, and evaluating claims; (b) nature of the learner, including intellectual, emotional-social, and physical development; the teaching-learning process, including kinds of learning, basic concepts and principles of learning, guidance of learning in the classroom; (c) sociological considerations, such as the influence of social class stratification on teaching, styles of teaching and patterns of social control, and the teacher's role as a member of a social system.
5. Evaluation and research appraisal, including (a) the justification and meaning of research conclusions and the bearing of evidence on educational decisions; current trends and issues and their historical perspective; (b) elementary statistical, measurement, and evaluation concepts and techniques bearing on the appraisal of methods, individuals, small groups and the broader society.

Students are called upon to demonstrate their knowledge of facts, terminology, theory and concepts, evidence, and professional sources at the same time they demonstrate their skill in using the cognitive processes—that is, recall of knowledge, comprehension, application, analysis, and evaluation. A typical question, for example, might require the student to make a prediction based on understanding of the social structure of the ghetto family. Another example would be a question that considers the justification of a teacher's particular course of action by the use of an appropriate generalization of a teaching theory.

Responses to Background Questions, 1970–71 (N = 24,179)

A. At what point are you in your studies?

1% (1) I am in or have just completed my junior year of undergraduate study.
24% (2) I am an undergraduate senior.
30% (3) I have a bachelor's degree but am not presently enrolled in graduate school.
28% (4) I am in or have just completed my first year of graduate study.
14% (5) I am in or have completed my second year of graduate study.

B. What graduate degree do you intend to seek?

2% (1) I do not plan to pursue graduate study.
3% (2) I plan to pursue graduate work but not to obtain a graduate degree.
69% (3) I plan to obtain terminal M.A., M.S., or other degree at the master's level.
13% (4) I plan to obtain M.A., M.S., or other master's level degree leading to a doctoral degree.
13% (5) I plan to obtain Ph.D., Ed.D., or other degree at the doctoral level.

C. In which of the following areas did you major as an undergraduate?

69% (1) Education (including elementary, secondary, and any related subject area specialization)
4% (2) Natural sciences, mathematics
11% (3) Social sciences
7% (4) Humanities, fine arts
7% (5) Other

D. In which of the following types of institutions did you do most of your undergraduate work?

26% (1) Four- or five-year college offering primarily liberal arts
22% (2) Four- or five-year college offering primarily teacher preparation
37% (3) State university
11% (4) Privately endowed university
3% (5) Other

E. If you are presently working toward a graduate degree in education, in which of the following areas are you concentrating your work?

14% (1) Administration and/or supervision
24% (2) Curriculum and instruction
5% (3) Psychological foundations (including educational psychology, human growth and development, mental hygiene, etc.)
2% (4) Social foundations (including history, philosophy, and sociology of education)
17% (5) Pupil personnel services (including guidance, special education, etc.)

F. How many semester (quarter) courses have you completed in the area of administration and/or supervision?

69% (1) None
7% (2) One
4% (3) Two
3% (4) Three
8% (5) Four or more

G. How many semester (quarter) courses have you completed in the area of curriculum and instruction?

31% (1) None
14% (2) One
11% (3) Two
8% (4) Three
27% (5) Four or more

H. How many semester (quarter) courses have you completed in the area of psychological foundations?

31% (1) None
19% (2) One
15% (3) Two
10% (4) Three
14% (5) Four or more

I. How many semester (quarter) courses have you completed in the area of social foundations?

43% (1) None
21% (2) One
10% (3) Two
6% (4) Three
9% (5) Four or more

Validity Data

1. The subjects of a study by Eckhoff (1966) were 185 secondary education majors and 111 elementary education majors with 30 or more quarter hours accumulated at Winona State College. The predictors were scores on the GRE Advanced Education Test and the Miller Anologies Test and undergraduate grade-point average. The criterion was overall graduate grade-point average. Stepwise regression analysis showed that the GRE scores added practically nothing to the prediction for secondary majors, and the MAT scores added nothing to the prediction for elementary majors. Elimination of these predictors then yielded the beta weights and multiple correlation coefficients between the two predictors and the criterion shown in Table 6.

Table 6: Beta Weights and Multiple Validity Coefficients Using Graduate Grades as the Criterion for Education Majors at Winona State College

	Beta Weights			Multiple Correlation
	GRE	MAT	UGPA	
Elementary (n = 111)	.18[1]	—	.21[2]	.30
Secondary (n = 185)	—	.23[2]	.41[2]	.51

[1] Significant at the .05 level
[2] Significant at the .01 level

2. The group involved in a study by Roscoe and Houston (1969) included 231 students who successfully completed the doctoral program in education at Colorado State College and an additional 21 students who were admitted, completed 30 quarter hours, and then were dismissed. The predictors included scores on the GRE Advanced Education Test and the GRE Aptitude Test (verbal and quantitative). The criteria were grade-point average in doctoral studies, graduation versus dismissal from the program, a normative judgment, and an ipsative judgment. To secure the normative judgments, test scores and other predictor variables on 30 representative students not identified by name were presented to 16 graduate professors. The professors rated the students' prospects as doctoral students. To get the ipsative judgments, the same 16

professors were given the names of doctoral graduates. The professors rated the professional promise of 10 students on the list whom they knew. The results are shown in Table 7.

Table 7: Validities for 252 Students in Education at Colorado State College

Predictors	Criteria			
	Grade-Point Average	Graduation vs Dismissal	Normative Judgment	Ipsative Judgment
GRE Advanced Education	.28	.26	.17	.30
GRE Verbal Ability	.32	.21	.38	.26
GRE Quantitative Ability	.21	.28	.27	.17

All correlations significant beyond the .01 level.

3. In a third study, by Williams, Harlow, and Grab (1970), the subjects were 84 students admitted to the doctoral program in the Education Department of the University of North Dakota between June 1962 and June 1967. The predictors included scores on the GRE Advanced Education Test, the GRE Aptitude Test, and the Miller Analogies Test, grade-point average for the bachelor's degree, and grade-point average for the master's degree. The criteria were doctoral grade-point average and graduation vs nongraduation. As of February 1969, 33 of the students had graduated and 51 had not graduated or been in attendance during the preceding 21 months. The correlations between predictors and criteria are shown in Table 8. Table 8A gives the means and standard deviations of the predictors and criteria.

Table 8: Validities for 84 Students in Education at the University of North Dakota

Predictors	Criteria	
	Doctoral Grades	Graduation vs Nongraduation
GRE Advanced Education	.08	.34[2]
GRE Verbal Ability	−.01	.08
GRE Quantitative Ability	−.01	.34[2]
MAT	−.03	.10
Grades for Bachelor's Degree	.13	.02
Grades for Master's Degree	−.01	.22[1]

[1] Significant at the .05 level

[2] Significant at the .01 level

Table 8A: Means and Standard Deviations of Predictors and Criterion for Students Graduated and Not Graduated

Predictor	Criterion	Graduated (n = 33)		Not Graduated (n = 51)	
		M	SD	M	SD
GRE Advanced Education		589	58	554	38
GRE Verbal Ability		515	75	503	71
GRE Quantitative Ability		586	101	504	71
MAT		51.7	8.54	49.5	11.4
Grades for Bachelor's Degree		2.75	.39	2.73	.39
Grades for Master's Degree		3.54	.24	3.43	.25
	Doctoral Grades	3.73	.31	3.51	.75

In addition to the six predictors listed above, nine other predictors were used. The 15 predictors gave a multiple correlation of .51 for doctoral grades and .65 for graduation vs nongraduation. The last multiple correlation was significant at the .01 level.

ADVANCED ENGINEERING TEST

Engineering is an extremely broad discipline. The subdisciplines of chemical, civil, electrical, industrial, and mechanical engineering are all represented on the committee of examiners and are included in the test. The aim is to ask engineering questions that are sufficiently fundamental and general so that all engineers, regardless of their specialty, can reasonably be expected to answer them. Since mathematics is basic to all branches of engineering, a substantial number of questions are devoted to mathematics. Two subscores, engineering and mathematics usage, are provided.

Content

Questions for the engineering subscore are based on material common to the several branches of engineering and usually studied during the first two collegiate years. The areas from which questions may be drawn are as follows:

Mechanics: statics, dynamics, kinematics, strength of materials; thermodynamics; fluid mechanics; transfer and rate mechanisms: heat, mass, momentum; structure of matter; electricity; chemistry; nature and properties of matter, including the particulate; light and sound; computer fundamentals; engineering judgement.

Approximately 80 to 90 questions on these topics make up this part of the examination.

Questions for the mathematics usage subscore have been developed from two viewpoints:

1. There is a body of intuitive mathematical concepts—in contrast to facts and formulas—that forms the basis upon which persons select from several possible approaches that one best fitted to a particular situation they have encountered in their discipline.
2. There is a body of knowledge of mathematical facts that should be at the fingertips of those who use mathematics—facts for which they cannot always seek verification if they are to work efficiently in their discipline.

The mathematics background assumed is not more than two courses in calculus with some simple ideas from linear algebra and probability—ideas that usually precede or accompany introductory calculus. This subscore is based on the following kinds of questions: intuitive calculus problems, factual recall questions, and a limited number of other types of mathematics questions.

The intuitive calculus questions consist of three sets of questions, approximately eight in each set, involving graphs of functions. In each set the students must use basic calculus ideas to interpret the given graphs and derive information concerning graphs or paths that are not drawn and that they construct for themselves from information available in the given figures in order to answer the questions.

In addition to the questions in the two mathematics usage parts, a few mathematics questions appear in the engineering parts of the test. Scores on these are part of the mathematics usage subscore.

Responses to Background Questions, 1970–71 (N = 7,858)

A. At what point are you in your studies?

3% (1) I am in or have just completed my junior year of undergraduate study.
52% (2) I am an undergraduate senior.
23% (3) I have a bachelor's degree but am not presently enrolled in graduate school.
11% (4) I am in or have just completed my first year of graduate study.
6% (5) I am in or have completed my second year of graduate study.

B. What graduate degree do you intend to seek?

5% (1) I do not plan to pursue graduate study.
2% (2) I plan to pursue graduate work but not to obtain a graduate degree.
52% (3) I plan to obtain terminal M.A., M.S., or other degree at the master's level.
25% (4) I plan to obtain M.A., M.S., or other master's level degree leading to a doctoral degree.
14% (5) I plan to obtain Ph.D., Ed.D., or other degree at the doctoral level.

What is the branch of engineering in which you are presently registered or were most recently registered? (Mark one space for question C or D to answer this question according to the following code.)

C.

12% (1) Chemical engineering
12% (2) Civil engineering
34% (3) Electrical engineering
3% (4) Industrial engineering
20% (5) Mechanical engineering

D.

1% (1) Agricultural engineering
6% (2) Aeronautical engineering
2% (3) Metallurgical engineering
1% (4) Nuclear engineering
9% (5) None of the above in either (C) or (D)

Validity Data

The subjects for a study by Creager (1965) were 300 male applicants for National Science Foundation fellowships in 1955 and 1956. The predictors were the GRE Aptitude Test (verbal and quantitative) and the Advanced Engineering Test. One criterion was time lapse between attainment of the B.A. and the Ph.D., coded as shown below:

B.A.-Ph.D. Time Lapse (in years):	Less than 4	4	5	6	7	8	9	No Ph.D. by Aug. '64
Coded Variable:	1	2	3	4	5	6	7	8

A second criterion was the dichotomous variable of attaining or not attaining a Ph.D. by August 1964. The third criterion was the dichotomous variable of attaining or not attaining a Ph.D. in the average time taken to attain a Ph.D. in the field. The relationships between predictors and criteria are shown in Table 9.

Table 9: Validities of GRE against Doctorate Attainment for 300 Males Who Were Applicants for National Science Foundation Fellowships in Engineering in 1955 and 1956

	Criteria				
		Ph.D. by 1964		Ph.D. in Average Time	
Predictors	Reflected B.A.-Ph.D. Time Lapse[1]	Point Biserial	Biserial	Point Biserial	Biserial
GRE Verbal Ability	.28	.28	.41	.28	.42
GRE Quantitative Ability	.21	.21	.31	.19	.28
GRE Advanced Engineering	.32	.31	.45	.31	.46
Composite	.35	.34	.50	.34	.50

[1] Correlations between the coded variable for B.A.-Ph.D. time lapse given above and the predictors with the signs reversed

ADVANCED FRENCH TEST

Content

Rather than stress any one content area, the committee responsible for the test tries to achieve a balanced approach. The test encompasses questions on reading comprehension, literary interpretation and criticism, literary history, and culture and civilization.

The sections on reading comprehension and on literary interpretation and criticism are designed to test comprehension on a variety of levels. Questions deal with vocabulary recognition and use of context to determine meaning, sensitivity to style and literary values, and ability to follow the development of an author's thought. Prose and poetry selections represent various periods and genres from the sixteenth century to the present. Texts vary considerably in length as well as in the number of questions based on them.

In the section on literary history, which includes the Middle Ages, questions require specific information on major works, authors, trends and movements and the ability to grasp significant relationships. Questions about French culture and civilization touch upon such topics as geography, history, institutions, and the arts and sciences. Other questions concern definitions of the genres, their evaluation, the vocabulary of rhetoric, and the ideas propounded in the last 30 years by the new critics, novelists, and dramatists. In this way, the committee attempts to recognize the critical approaches to literature that are becoming more prevalent without neglecting the broad evolutionary perspectives still considered valid.

Literary selections and questions give approximately equal attention to all centuries from the sixteenth through the twentieth. Knowledge of the language (grammar, idioms) is *not* tested in questions separate from interpretive reading selections, and linguistics is not tested at all.

Responses to Background Questions, 1970–71
(N = 2,472)

A. At what point are you in your studies?

3% (1) I am in or have just completed my junior year of undergraduate study.
65% (2) I am an undergraduate senior.
20% (3) I have a bachelor's degree but am not presently enrolled in graduate school.
7% (4) I am in or have just completed my first year of graduate study.
4% (5) I am in or have completed my second year of graduate study.

B. What graduate degree do you intend to seek?

4% (1) I do not plan to pursue graduate study.
3% (2) I plan to pursue graduate work but not to obtain a graduate degree.
49% (3) I plan to obtain terminal M.A., M.S., or other degree at the master's level.
26% (4) I plan to obtain M.A., M.S., or other master's level degree leading to a doctoral degree.
17% (5) I plan to obtain Ph.D., Ed.D., or other degree at the doctoral level.

C. Was French regularly spoken in your home when you were a child?

9% (1) Yes
90% (2) No

D. For what length of time have you studied in or lived in a French-speaking country?

33% (1) Not at all
21% (2) Some, but less than three months
8% (3) Three to six months
23% (4) Six months to one year
15% (5) More than one year

E. What is (or was) your undergraduate major field?

85% (1) French
3% (2) Another foreign language
11% (3) Other

F. If you majored in French as an undergraduate, which of the following was most emphasized in your courses?

65% (1) French literature
17% (2) French language proficiency
2% (3) Civilization and culture (including area studies)
2% (4) Linguistics (history of language, structure of language)
3% (5) Other

ADVANCED GEOGRAPHY TEST

Content

The questions in the Advanced Geography Test are drawn from the courses of study most commonly offered. Approximately 40 percent of the questions are devoted to physical geography and 60 percent to human geography. Questions on physical geography deal with such topics as climate, landforms, biogeography, vegetation, soils, the environment as a system, water, and cartography. Questions on human geography cover such areas as economic geography, resources, transportation, trade, settlement, and population. Some questions require knowledge of more than one area or more than one aspect of geography. Increasing emphasis is being placed by the committee responsible for the test on including questions that call for application by the students of their powers of reasoning and analytical skills, rather than merely their capacity for recall.

Responses to Background Questions, 1970–71 (N = 962)

A. At what point are you in your studies?

2% (1) I am in or have just completed my junior year of undergraduate study.
53% (2) I am an undergraduate senior.
21% (3) I have a bachelor's degree but am not presently enrolled in graduate school.
12% (4) I am in or have just completed my first year of graduate study.
10% (5) I am in or have completed my second year of graduate study.

B. What graduate degree do you intend to seek?

1% (1) I do not plan to pursue graduate study.
1% (2) I plan to pursue graduate work but not to obtain a graduate degree.
47% (3) I plan to obtain terminal M.A., M.S., or other degree at the master's level.
35% (4) I plan to obtain M.A., M.S., or other master's level degree leading to a doctoral degree.
14% (5) I plan to obtain Ph.D., Ed.D., or other degree at the doctoral level.

C. How many semester (quarter) courses have you had in physical geography?

30% (1) One or fewer
24% (2) Two
22% (3) Three or four
10% (4) Five or six
12% (5) Seven or more

D. How many semester (quarter) courses have you had in economic geography?

66% (1) One or fewer
18% (2) Two
8% (3) Three or four
4% (4) Five or six
2% (5) Seven or more

E. How many semester (quarter) courses have you had in cultural geography?

41% (1) One or fewer
21% (2) Two
19% (3) Three or four
10% (4) Five or six
7% (5) Seven or more

F. Have you had any courses in geographic thought and methods?

53% (1) Yes
46% (2) No

G. Have you had any courses in cartography?

63% (1) Yes
35% (2) No

ADVANCED GEOLOGY TEST

Content

Modern geological thinking crosses many subject boundaries, and numerous questions in the test reflect this tendency. Nevertheless, each question reasonably falls into one of three major categories.

A separate subscore is reported for each of these three categories. A further description of the content follows.

I. STRATIGRAPHY, PALEONTOLOGY, AND GEOMORPHOLOGY — 70 questions
  - A. Stratigraphy
  - B. Sedimentology
  - C. Paleontology (invertebrate and vertebrate)
  - D. History
  - E. Geomorphology, including glaciology
  - F. General, including oceanography

II. STRUCTURAL GEOLOGY AND GEOPHYSICS — 70 questions
  - A. Structure—field relations
  - B. Structure—dynamics (experimental and theoretical)
  - C. Tectonics
  - D. Isostasy, gravity, and magnetism
  - E. Earthquakes and seismology
  - F. Heat and electrical properties
  - G. General, including planetology

III. MINERALOGY, PETROLOGY, AND GEOCHEMISTRY — 60 questions
  - A. Mineralogy
    - 1. Chemical compositions
    - 2. Physical properties (optical, x-rays, and crystallography)
  - B. Petrology
    - 1. Field relations
    - 2. Compositions and mineral assemblages of rocks
  - C. Geochemistry
    - 1. Solutions
    - 2. Phase equilibria
  - D. Radiometric dating
  - E. Economic mineral deposits

The Advanced Geology Test is designed to measure important abilities, as follows:

- Ability to analyze geologic phenomena using, for example, maps, graphs, cross sections, thin sections, block diagrams, diagrams resulting from instrumental methods, and perceptions in three dimensions.
- Ability to comprehend geological processes, including comprehension through the application of physics, chemistry, biology, and mathematics.
- Ability to demonstrate knowledge of basic geology.

Responses to Background Questions, 1970–71
(N = 1,636)

A. At what point are you in your studies?

3% (1) I am in or have just completed my junior year of undergraduate study.
64% (2) I am an undergraduate senior.
15% (3) I have a bachelor's degree but am not presently enrolled in graduate school.
9% (4) I am in or have just completed my first year of graduate study.
8% (5) I am in or have completed my second year of graduate study.

B. What graduate degree do you intend to seek?

2% (1) I do not plan to pursue graduate study.
1% (2) I plan to pursue graduate work but not to obtain a graduate degree.
37% (3) I plan to obtain terminal M.A., M.S., or other degree at the master's level.
37% (4) I plan to obtain M.A., M.S., or other master's level degree leading to a doctoral degree.
22% (5) I plan to obtain Ph.D., Ed.D., or other degree at the doctoral level.

C. Did you major in geology as an undergraduate?

87% (1) Yes
11% (2) No

D. If geology was *not* your undergraduate major, in which of the following fields did you concentrate as an undergraduate?

2% (1) Biology
1% (2) Chemistry
4% (3) Physics
2% (4) Mathematics
9% (5) Other

E. With respect to graduate schools, why are you taking this test?

26% (1) To gain admission to a graduate school only.
7% (2) To secure financial assistance from a graduate school only.
57% (3) To gain admission to and secure financial assistance from a graduate school.

F. Are you taking this test in order to secure financial assistance from the National Science Foundation?

16% (1) Yes
78% (2) No

Validity Data

1. A group reported on by the Office of Educational Research (1968) was composed of 78 students registering for graduate work in geology at Yale University in the years 1952–1961, inclusive. The predictors were scores on the GRE Advanced Geology Test and the GRE Aptitude Test (verbal and quantitative). The criterion was a

composite rating of the students by faculty members. The correlations with this criterion were .51 for Advanced Geology, .32 for verbal ability, .38 for quantitative ability, and .54 for an optimally weighted composite of the three predictors.

2. The subjects for a study by Creager (1965) were 119 male applicants for National Science Foundation fellowships in 1955 and 1956. The predictors were scores on the GRE Aptitude Test (verbal and quantitative) and the Advanced Geology Test. One criterion was time lapse between attainment of the B.A. and the Ph.D., coded as shown below:

B.A.-Ph.D. Time Lapse (in years):	Less than 4	4	5	6	7	8	9	No Ph.D. by Aug. '64
Coded Variable:	1	2	3	4	5	6	7	8

A second criterion was the dichotomous variable of attaining or not attaining a Ph.D. by August 1964. The third criterion was the dichotomous variable of attaining or not attaining a Ph.D. in the average time taken to attain a Ph.D. in the field. The relationships between predictors and criteria are shown in Table 10.

Table 10: Validities of GRE against Doctorate Attainment for 119 Males Who Were Applicants for National Science Foundation Fellowships in Geology in 1955 and 1956

	Criteria				
		Ph.D. by 1964		Ph.D. in Average Time	
Predictors	Reflected B.A.-Ph.D. Time Lapse[1]	Point Biserial	Biserial	Point Biserial	Biserial
GRE Verbal Ability	.25	.30	.41	.26	.36
GRE Quantitative Ability	.26	.27	.37	.24	.33
GRE Advanced Geology	.22	.20	.27	.22	.30
Composite	.31	.33	.45	.32	.44

[1] Correlations between the coded variable for B.A.-Ph.D. time lapse given above and the predictors with the signs reversed

ADVANCED GERMAN TEST

Content

Rather than stress any one content area, the committee responsible for the test endeavors to achieve a balanced approach. The test encompasses questions on German structure and idiomatic usage, reading comprehension, literary interpretation and sensitivity, literary history, and culture and civilization. A few questions also touch on basic concepts of linguistics.

Prose and poetry selections, principally from nineteenth-century and twentieth-century literature and of various degrees of difficulty, are designed to test comprehension on a variety of levels. In addition to measuring accurate comprehension of content, questions deal with literary expression, sensitivity to style and literary values, literary criticism, and the ability to follow the development of an author's thought. The questions on literary history, from the Middle Ages to the present, require specific information on major works, authors, trends, and movements and the ability to grasp significant relationships. Questions on German culture and civilization touch upon history, geography, institutions, science, and the arts.

Literary selections stress the twentieth century heavily and none are earlier than the nineteenth century; both fiction and nonfiction are represented. However, questions on literary facts embrace the entire history of German literature from the Middle Ages on. Sensitivity to literary style is tested through specific item types. Knowledge of grammar and idioms and applied linguistics are tested separately.

Responses to Background Questions, 1970–71
(N = 702)

A. At what point are you in your studies?

3%	(1) I am in or have just completed my junior year of undergraduate study.
64%	(2) I am an undergraduate senior.
18%	(3) I have a bachelor's degree but am not presently enrolled in graduate school.
7%	(4) I am in or have just completed my first year of graduate study.
6%	(5) I am in or have just completed my second year of graduate study.

B. What graduate degree do you intend to seek?

3%	(1) I do not plan to pursue graduate study.
3%	(2) I plan to pursue graduate work but not to obtain a graduate degree.
39%	(3) I plan to obtain terminal M.A., M.S., or other degree at the master's level.
27%	(4) I plan to obtain M.A., M.S., or other master's level degree leading to a doctoral degree.
26%	(5) I plan to obtain Ph.D., Ed.D. or other degree at the doctoral level.

C. Was German regularly spoken in your home when you were a child?

17%	(1) Yes
82%	(2) No

D. Do you now speak German with native or near-native fluency?

51%	(1) Yes
47%	(2) No

E. When did you begin to study German?

10%	(1) In grade school
7%	(2) In junior high school
42%	(3) In high school
27%	(4) As a college freshman
12%	(5) As a college sophomore or later

F. For what length of time have you studied in or lived in a German-speaking country?

21%	(1) Not at all
14%	(2) Some, but less than three months
9%	(3) Three to six months
25%	(4) Six months to one year
30%	(5) More than one year

G. What is (or was) your undergraduate major field?

79%	(1) German
4%	(2) Another foreign language
17%	(3) Other

H. If you majored in German as an undergraduate, which of the following was most emphasized in your courses?

59%	(1) Literature
14%	(2) Language proficiency
3%	(3) Civilization and culture (including area studies)
3%	(4) Linguistics (history of language, structure of language)
4%	(5) Other

ADVANCED HISTORY TEST

Content

The questions in the test are drawn from the courses of study most commonly offered. Major considerations that have determined the content and form of the Advanced History Test are the uses made of the test and the wide variation in preparation of the students taking it.

In other words, the test provides one measure of the experience an undergraduate major in history has acquired in the discipline of history and of the knowledge and abilities required for graduate work in history. More specifically, this experience consists of (1) familiarity with historical data and (2) the ability to apply knowledge gained by this familiarity, particularly to perceive relationships—both those involving individuals and movements and those that are chronological—and to analyze historical material in various forms, such as historical documents or passages from historical works. The potential graduate student should also have an understanding of the meaning and use of sources and the significance of movements and periods. Thus, the test measures factual knowledge not for itself but as it facilitates the understanding of periods, trends, and relationships.

The problem of content coverage in a single history test is complex. It is almost impossible to delimit the field of history in area, in time, and in scope. Moreover, no common core of knowledge is required of all history majors in all colleges. In the Advanced History Test, all the questions refer to the history of the United States and Europe (somewhat more questions are devoted to the latter than the former) because these remain the areas in which the greatest number of students concentrate. There may be questions involving an understanding of the relationships among these and other areas. Similarly, questions deal with economic social, and intellectual—as well as political—history in about equal proportions. Although the majority of the questions concern the period after 1789, the Middle Ages, the Renaissance, and the Reformation are also covered.

The committee responsible for the test is acutely aware that histories other than those of Europe and the United States have increasingly assumed a greater place in the curriculum, both as required courses and electives. For many years, in fact, the GRE Advanced History Test included questions on Asian, African, and Latin American history. In 1969 the committe reluctantly decided to stop including such questions in the test. Given the limited number of students enrolled in courses in other than European and United States history, it had always been difficult to judge how many questions could provide adequate coverage or useful conclusions. What was decisive, however, was the evidence in the test results that the examinees as a whole did poorly on such questions. Although the committee would like to have kept them in the test as acknowledgment of the rising importance of Asian, African, and Latin American studies, it was neither fair to the majority of students nor good test-making practice to present questions that—however basic—were unlikely to be answered correctly by able students with extensive backgrounds in the prevailing European and American history curriculums.

More recently, on the basis of a course survey available to the committee from a background questionnaire filled out by examinees in Advanced History, it was decided—again reluctantly—not to resume testing in Asian, African, and Latin American history at this time. The committee, however, continues to monitor curriculum developments to determine whether a change in test content is warranted.

History is typical of most of the Advanced Tests in that it is a highly reliable test but one of greater than middle difficulty for the test population. Attempts to reduce the difficulty of the History Test are being made through decreasing the reading time needed to answer the application questions, diminishing the emphasis on economic history, and posing more basic questions on Russian and Eastern European history.

Responses to Background Questions, 1970–71 (N = 10,637)

A. At what point are you in your studies?

3% (1) I am in or have just completed my junior year of undergraduate study.
62% (2) I am an undergraduate senior
21% (3) I have a bachelor's degree but am not presently enrolled in graduate school.
9% (4) I am in or have just completed my first year of graduate study.
4% (5) I am in or have completed my second year of graduate study.

B. What graduate degree do you intend to seek?

4% (1) I do not plan to pursue graduate study.
2% (2) I plan to pursue graduate work but not to obtain a graduate degree.
40% (3) I plan to obtain terminal M.A., M.S., or other degree at the master's level.
31% (4) I plan to obtain M.A., M.S., or other master's level degree leading to a doctoral degree.
20% (5) I plan to obtain Ph.D., Ed.D., or other degree at the doctoral level.

In answering questions C through I, include the courses you are presently taking.

C. How many semester (quarter) courses have you had in United States history (including the Colonial period)?

5% (1) None
25% (2) One or two
34% (3) Three or four
21% (4) Five or six
15% (5) Seven or more

D. How many semester (quarter) courses have you had in ancient history and medieval European history (including courses in the history of individual European countries)?

19% (1) None
27% (2) One
24% (3) Two
13% (4) Three
16% (5) Four or more

E. How many semester (quarter) courses have you had in the Renaissance and early modern European history to 1789 (including courses in the history of individual European countries)?

24% (1) None
37% (2) One
23% (3) Two
9% (4) Three
6% (5) Four or more

F. How many semester (quarter) courses have you had in modern European history from 1789 to the present (including courses in the history of individual European countries)?

16% (1) None
29% (2) One
24% (3) Two
14% (4) Three
16% (5) Four or more

G. How many semester (quarter) courses have you had in the history of China and/or Japan?

68% (1) None
20% (2) One
7% (3) Two
2% (4) Three
2% (5) Four or more

H. How many semester (quarter) courses have you had in the history of Africa (including courses in the history of individual African countries)?

84% (1) None
11% (2) One
2% (3) Two
1% (4) Three
1% (5) Four or more

I. How many semester (quarter) courses have you had in the history of Latin America (including courses in the history of individual Latin American countries?

73% (1) None
17% (2) One
6% (3) Two
2% (4) Three
2% (5) Four or more

Validity Data

1. A group reported on by Johnson and Thompson (1962) was composed of a small number of graduate students in history at Sacramento State College. The correlation between the predictor of GRE Advanced History Test scores and the criterion of grade-point averages in all graduate study was .56. The coefficient was significant at the .05 level.

2. A group studied by Lannholm, Marco, and Schrader (1968) was composed of 66 students first enrolled in a particular graduate department of history between the fall of 1957 and June 1960. The predictors were scores on the GRE Advanced History Test and the GRE Aptitude Test, and undergraduate grade-point average. The criterion was success in graduate study, defined as having earned the Ph.D. or being still enrolled and rated by faculty members as outstanding or superior in the fall of 1963. The results are shown in Table 11.

Table 11: Validities for 66 History Students

Predictor	r-biserial Correlation with Criterion	Mean Performance on Predictor	Standard Deviation of Performance on Predictor
GRE Advanced History	.51	565	98
GRE Verbal Ability	.41	575	115
GRE Quantitative Ability	.36	509	128
UGPA	.44	2.93	.56
Optimally Weighted Combination	.59		

3. Another group studied by Lannholm et al. (1968) was composed of 28 students first enrolled in a particular graduate department of history between the fall of 1957 and June 1960. The predictors were scores on the GRE Advanced History Test and the GRE Aptitude Test. The criterion was success in graduate study, defined as having earned the Ph.D. or being still enrolled and rated by faculty members as outstanding or superior in the fall of 1963. The results are shown in Table 12.

Table 12: Validities for 28 History Students

Predictor	r-biserial Correlation with Criterion	Mean Performance on Predictor	Standard Deviation of Performance on Predictor
GRE Advanced History	.46	596	77
GRE Verbal Ability	–.04	647	84
GRE Quantitative Ability	–.13	549	101

4. Roberts (1970) studied the records of 63 students who had enrolled at Wake Forest University from June 1964 to June 1970 for graduate study in history and who had completed at least nine hours of graduate work. The correlations between graduate grade-point averages and GRE scores were –.31 for verbal ability, –.18 for quantitative ability, and –.31 for Advanced History.

ADVANCED LITERATURE IN ENGLISH TEST

Content

The test contains questions on poetry, drama, biography, the essay, criticism, the short story, the novel, and, to a limited extent, the history of the language. The test draws on English and American literature of all periods; it also contains a few questions on well-known foreign writers and on works, including the Bible, translated from foreign languages. Throughout, the emphasis is on major authors, works, and movements.

The questions may be somewhat arbitrarily classified into two groups: factual and critical. The factual questions test a student's knowledge of the major writers usually studied in college literature courses. For example, the student may be asked to identify a writer's major contribution to literary history, to assign a literary work to the period in which it was written, to identify the primary theme of a work, to identify common kinds of poetic meter, to recognize a literary allusion in a given context, to identify a writer or work described in a brief critical comment, or to determine the period or author of a work on the basis of the style and content of a short excerpt. The critical questions test the ability to read a literary text perceptively. Students are asked to examine a given passage of prose or poetry and to answer questions about the author's thesis or ideas and his or her use of figurative language. Such questions also deal with form and structure, literary techniques, and various aspects of style.

Often examinees will feel the test has discovered and emphasized areas in which they are least adequate. In fact examinees tend to remember most vividly those questions that proved troublesome. Students taking the GRE should remember that, in a test of approximately 230 questions, much of the material presented no undue difficulty. The very length and scope of the examination eventually work to the benefit of the students and give them an opportunity to demonstrate what they do know. No one is expected to answer all the questions correctly.

Responses to Background Questions, 1970–71
(N = 14,079)

A. At what point are you in your studies?

2% (1) I am in or have just completed my junior year of undergraduate study.
60% (2) I am an undergraduate senior.
22% (3) I have a bachelor's degree but am not presently enrolled in graduate school.
10% (4) I am in or have just completed my first year of graduate study.
5% (5) I am in or have completed my second year of graduate study.

B. What graduate degree do you intend to seek?

3% (1) I do not plan to pursue graduate study.
2% (2) I plan to pursue graduate work but not to obtain a graduate degree.
43% (3) I plan to obtain terminal M.A., M.S., or other degree at the master's level.
30% (4) I plan to obtain M.A., M.S., or other master's level degree leading to a doctoral degree.
20% (5) I plan to obtain Ph.D., Ed.D., or other degree at the doctoral level.

C. What was your undergraduate major?

90% (1) English
2% (2) History or philosophy
2% (3) Social science
1% (4) Foreign language
1% (5) A natural science or mathematics

D. If you studied English in college, in which of the following areas did you concentrate?

59% (1) English literature
17% (2) American literature
10% (3) Comparative literature
1% (4) Linguistics
4% (5) Composition and rhetoric

E. How many semester (quarter) courses in English literature have you taken?

2% (1) Fewer than two
13% (2) Two or three
22% (3) Four or five
21% (4) Six or seven
42% (5) Eight or more

F. How many semester (quarter) courses in American literature have you taken?

27% (1) Fewer than two
45% (2) Two or three
18% (3) Four or five
6% (4) Six or seven
4% (5) Eight or more

G. Which of the following best describes a comprehensive historical survey of English literature (as opposed to a period or genre course) which you may have taken in college?

54% (1) A two-semester full survey (at least from Chaucer to the twentieth century)
10% (2) A one-semester full survey
6% (3) The first half of a two-semester survey
3% (4) The second half of a two-semester survey
25% (5) No survey course taken

H. In what period has your undergraduate preparation been most thorough (in number of hours taken)?

5% (1) Old and Middle English
14% (2) Renaissance
10% (3) Restoration and eighteenth century
26% (4) Romantic and Victorian
38% (5) Modern

I. In what genre has your undergraduate preparation been most thorough (in number of hours taken)?

10% (1) Drama
31% (2) Poetry
47% (3) Fiction
4% (4) Prose nonfiction
4% (5) Other

Validity Data

1. The group included in a study by Lannholm, Marco, and Schrader (1968) was composed of 98 students first enrolled in a particular graduate department of English between the fall of 1957 and June 1960. The predictors were scores on the GRE Advanced Literature in English Test and the GRE Aptitude Test. The criterion was success in graduate study, defined as having earned the Ph.D. or being still enrolled and rated by faculty members as outstanding or superior in the fall of 1962. The results are reported in Table 13.

Table 13: Validities for 98 Literature Students

Predictor	r-biserial Correlation with Criterion	Mean Performance on Predictor	Standard Deviation of Performance on Predictor
GRE Advanced Literature in English	.31	705	58
GRE Verbal Ability	.25	724	59
GRE Quantitative Ability	.13	576	106
Optimally Weighted Combination	.34		

2. The group in a second study by Lannholm et al. (1968) was composed of 81 students first enrolled in a particular graduate department of English between the fall of 1957 and June 1960. The predictors were scores on the GRE Advanced Literature in English Test and the GRE Aptitude Test and undergraduate grade-point average. The criterion was success in graduate study, defined as having earned the Ph.D. or being still enrolled and rated by faculty members as outstanding or superior in the fall of 1963. The results are reported in Table 14.

Table 14: Validities for 81 Literature Students

Predictor	r-biserial Correlation with Criterion	Mean Performance on Predictor	Standard Deviation of Performance on Predictor
GRE Advanced Literature in English	.43	626	87
GRE Verbal Ability	.32	611	102
GRE Quantitative Ability	.45	490	96
UGPA	.49	3.08	.51
Optimally Weighted Combination	.67		

3. Roberts (1970) studied the records of 60 students who had enrolled at Wake Forest University from June 1964 to June 1970 for graduate study in English and who had completed at least nine hours of graduate work. The correlations between graduate grade-point averages and GRE scores were .17 for verbal ability, .01 for quantitative ability, and .54 for Advanced Literature in English.

ADVANCED MATHEMATICS TEST

Content

The questions in the test are drawn from the courses of study most commonly offered. Approximately 50 percent of the questions involve analysis and its applications—subject matter that can be assumed to be common to the backgrounds of almost all mathematics majors. About 25 percent of the questions in the test are in linear and abstract algebra.

The remaining portion consists of a few questions in each of several other areas of mathematics currently offered to undergraduates in many institutions. Included are such areas as probability and statistics, number theory, set theory, logic, combinatorial analysis, topology, numerical analysis, and computer programming. There are also questions that ask the candidate to match "real-life" situations to appropriate mathematical models.

Because the material in the last-mentioned 25 percent of the test is so diverse, no useful purpose would be served in attempting to describe any substantial part of it. However, the material in analysis and algebra, on which 75 percent of the test is based, is probably well enough defined to make the following somewhat more detailed description useful.

Analysis. The usual material of two years of calculus, including trigonometry, analytic geometry (through conic sections and quadric surfaces), and introductory differential equations; introductory real variable theory, as presented in courses such as those entitled "advanced calculus" or "methods of real analysis" that include the elementary topology of the line, plane, 3-space, and n-space as well as Riemann integration, and, it is hoped, Stieltjes and Lebesgue integration.

Topics in Algebra. Group (abelian, cyclic), subgroup, normal subgroup, quotient group, permutation group, order (of group, of element), La Grange's Theorem, ring, ideal, integral domain, zero divisor, field, polynomial ring, congruence modulo an integer, divisibility, division algorithm (for integers, polynomials), homomorphism, isomorphism, and automorphism (for groups, rings, fields), vector space, kernel, null space, dimension, linear independence, dual space, inner product space, linear transformation, matrix of a linear transformation, characteristic root, trace, determinant, matric operations, similarity of matrices, spectral theorem for normal matrices (possibility of diagonalization).

Neither the description of analysis nor the list of topics in algebra is intended to be exhaustive. Obviously, it is necessary to understand many other concepts, but it is hoped the description will provide the prospective examinee with a useful idea of the material past and present committees of examiners have considered, and now consider, to be appropriate for a test designed to measure knowledge, skills, and aptitude needed for graduate study in mathematics.

It should be emphasized that *knowledge* of the material described above is a necessary, but not sufficient, condition for correctly answering the questions in analysis and algebra. Actually, a substantial number of questions require no more than a good precalculus background; and, in general, questions are intended to test more than straight knowledge and, indeed, concentrate on testing (1) understanding of fundamental concepts and (2) the ability to choose among and apply these concepts in novel situations.

A substantial number of "insightful" questions are included. Such questions have at least two avenues of approach, one obvious and requiring tedious manipulations, and the other not at all obvious but requiring little, if any, computation or manipulation.

Responses to Background Questions, 1970–71
(N = 7,131)

A. At what point are you in your studies?

4% (1) I am in or have just completed my junior year of undergraduate study.
64% (2) I am an undergraduate senior.
16% (3) I have a bachelor's degree but am not presently enrolled in graduate school.
9% (4) I am in or have just completed my first year of graduate study.
6% (5) I am in or have completed my second year of graduate study.

B. What graduate degree do you intend to seek?

5% (1) I do not plan to pursue graduate study.
2% (2) I plan to pursue graduate work but not to obtain a graduate degree.
38% (3) I plan to obtain terminal M.A., M.S., or other degree at the master's level.
24% (4) I plan to obtain M.A., M.S., or other master's level degree leading to a doctoral degree.
28% (5) I plan to obtain Ph.D., Ed.D., or other degree at the doctoral level.

C. Which of the following best describes your reason for taking this examination?

9% (1) It is required to qualify for a National Science Foundation Fellowship.
44% (2) It is required in applying for a graduate school fellowship or assistantship.
17% (3) It is required for continuing graduate study at my institution.
7% (4) It is required for earning an undergraduate degree at my institution.
21% (5) Other

D. How many semester (quarter) courses in mathematics have you taken above the level of precalculus mathematics?

24% (1) Eight or fewer
28% (2) Nine to eleven
21% (3) Twelve to fifteen
10% (4) Sixteen to twenty
15% (5) Twenty-one or more

E. In which of the following areas of mathematics has your study been most concentrated?

43% (1) Real analysis
3% (2) Complex analysis
37% (3) Algebra
3% (4) Topology
3% (5) Geometry

F. In what area other than mathematics was your undergraduate preparation strongest?

15% (1) Computer science
24% (2) Physics
11% (3) A natural science other than physics
6% (4) Philosophy
41% (5) Other

G. How many semester (quarter) courses have you taken in computer science?

36% (1) No undergraduate or graduate
44% (2) One or two undergraduate; no graduate
13% (3) More than two undergraduate; no graduate
4% (4) One or two graduate
2% (5) Three or more graduate

H. What is the total number of mathematics semester (quarter) courses in which you were tested primarily on your ability to write out proofs?

13% (1) None
14% (2) One
19% (3) Two
26% (4) Three or four
26% (5) Five or more

I. What is the total number of mathematics semester (quarter) courses in which you were tested primarily on your ability to solve problems in the sense of obtaining a numerical answer to a problem?

15% (1) Two or fewer
25% (2) Three or four
24% (3) Five or six
27% (4) Seven or more but not all
5% (5) All

Table 15: Validities Using the Criterion of Attainment of the Doctorate for 845 National Science Foundation Fellowship Applicants in Mathematics in 1958–61, Split into Two Random Halves

	n = 423			n = 422		
Predictors	r-biserial Correlation with Criterion	Predictor Performance Mean	Predictor Performance Standard Deviation	r-biserial Correlation with Criterion	Predictor Performance Mean	Predictor Performance Standard Deviation
GRE Advanced Mathematics[1]	.38	65.93	15.39	.44	64.93	15.94
GRE Verbal Ability[1]	.27	62.95	10.96	.32	62.63	11.33
GRE Quantitative Ability[1]	.27	72.67	9.51	.26	71.54	10.14
UGPA[2]	.21	252.60	40.22	.24	248.77	43.13
Reference Letters[3]	.23.	42.60	9.38	.27	42.59	9.69

[1] Scaled score with third digit dropped
[2] On a four-point scale multiplied by 100
[3] Zero to 6 multiplied by 10

Validity Data

1. A group studied by Rock (1972) was composed of 845 applicants for National Science Foundation fellowships. Most applied for first-year NSF fellowships in 1958–61. The predictors were scores on the GRE Advanced Mathematics Test and the GRE Aptitude Test, undergraduate grade-point average, and an average rating of reference letters. The criterion was attainment of the doctorate by June 1968. The group was split into random halves; the validity coefficients for each half are shown in Table 15.

2. The group involved in an earlier study by Johnson and Thompson (1962) was composed of a small number of graduate students in mathematics at Sacramento State College. The correlation between the predictor of GRE Advanced Mathematics Test score and the criterion of grade-point average for all graduate study was .76. This coefficient was significant at the .05 level.

3. Roberts (1970) studied the records of 37 students who had enrolled at Wake Forest University from June 1964 to June 1970 for graduate study in mathematics and who had completed at least nine hours of graduate work. The correlations between graduate grade-point averages and GRE scores were .41 for verbal ability, .55 for quantitative ability, and .47 for Advanced Mathematics.

4. The subjects for a study by Creager (1965) were 250 male applicants for National Science Foundation fellowships in 1955 and 1956. The predictors were scores on the GRE Aptitude Test (verbal and quantitative) and the Advanced Mathematics Test. One criterion was time lapse between attainment of the B.A. and the Ph.D., coded as shown below:

B.A.-Ph.D. Time Lapse (in years):	Less than 4	4	5	6	7	8	9	No. Ph.D. by Aug. '64
Coded Variable:	1	2	3	4	5	6	7	8

A second criterion was the dichotomous variable of attaining or not attaining a Ph.D. by August 1964. The third criterion was the dichotomous variable of attaining or not attaining a Ph.D. in the average time taken to attain a Ph.D. in the field. The relationships between predictors and criteria are shown in Table 16.

Table 16: Validities of GRE against Doctorate Attainment for 250 Males Who Were Applicants for National Science Foundation Fellowships in Mathematics in 1955 and 1956

	Criteria				
		Ph.D. by 1964		Ph.D. in Average Time	
Predictors	Reflected B.A.-Ph.D. Time Lapse[1]	Point Biserial	Biserial	Point Biserial	Biserial
GRE Verbal Ability	.21	.22	.30	.19	.27
GRE Quantitative Ability	.25	.26	.36	.23	.32
GRE Advanced Mathematics	.36	.34	.47	.34	.48
Composite	.36	.35	.48	.34	.48

[1] Correlations between the coded variable for B.A.-Ph.D. time lapse given above and the predictors with the signs reversed.

ADVANCED MUSIC TEST

Content

Questions in the test are drawn from the courses of study most commonly offered. Examination of curriculum offerings reveals that two major areas of study—music theory and music history—constitute the core of most undergraduate music programs. These two areas provide the content focus of the Advanced Music Test. About 40 percent of the questions deal with music theory; the remainder deal with music history and literature. Since a number of the questions relate to style analysis, which may combine aspects of both history and theory, the percentages are approximations only.

Approximately three-fourths of the theory questions deal with traditional theory and about one-fourth with contemporary theory. Aspects of theory covered in the test range from such fundamental concepts as scales, intervals, and key signatures to concepts related to jazz and contemporary composition. Other examples of topics in music theory include cadences, canon, fugue, modes, rhythmic devices, principles of instrumentation and orchestration, quartal harmony, polychords, serial music, and electronic music.

Questions dealing with music history and literature cover four historical periods—Medieval-Renaissance, Baroque, Classical-Romantic-Impressionistic, and Twentieth Century. The questions are relatively evenly divided among the four periods.

A number of cognitive abilities are measured in the test. Approximately one third of the questions in the test are devoted to style analysis. These questions are intended to test ability to interrelate facets of musical knowledge, such as styles, composers, and historical periods, and ability to analyze musical passages, including score reading—with application of appropriate principles of theory, harmony, and instrumentation. The remaining two-thirds of the questions measure familiarity with basic musical terminology, concepts, and principles; ability to read and interpret musical notation; and ability to identify, from written musical notation, compositions and such musical elements as intervals and scales.

Listening skills—which are not tested—are, of course, basic to the study of music. An experimental listening test was administered and its validity studied in the early 1960's. However, the limitations of a testing program that must accommodate the needs of a variety of students have prevented the inclusion of such a listening measure. It should be noted that correlations between the experimental listening test and the written test are quite high.

Responses to Background Questions, 1970–71
(N = 2,503)

A. At what point are you in your studies?

2% (1) I am in or have just completed my junior year of undergraduate study.
53% (2) I am an undergraduate senior.
24% (3) I have a bachelor's degree but am not presently enrolled in graduate school.
12% (4) I am in or have just completed my first year of graduate study.
8% (5) I am in or have completed my second year of graduate study.

B. What graduate degree do you intend to seek?

2% (1) I do not plan to pursue graduate study.
3% (2) I plan to pursue graduate work but not to obtain a graduate degree.
53% (3) I plan to obtain terminal M.A., M.S., or other degree at the master's level.
27% (4) I plan to obtain M.A., M.S., or other master's level degree leading to a doctoral degree.
13% (5) I plan to obtain Ph.D., Ed.D., or other degree at the doctoral level.

C. In which of the following areas is your music specialty?

43% (1) Music education
36% (2) Applied music (performance)
8% (3) Music history and literature-musicology
9% (4) Music theory-composition
4% (5) Other

D. If you have a declared major instrument, within which of the following group does it fall?

24% (1) Voice
37% (2) Piano-organ
6% (3) Strings
26% (4) Woodwinds-brass
2% (5) Percussion

E. Which of the following most accurately identifies the type of institution in which your training is currently being received?

21% (1) Music department in a teacher education institution
51% (2) Music department in a liberal arts institution
12% (3) Music school or fine arts school
7% (4) Music conservatory
5% (5) Other

F. How many semester (quarter) courses have you had specifically in ear training and sight singing?

22% (1) None
10% (2) One
18% (3) Two
12% (4) Three
35% (5) Four or more

G. How many semester (quarter) courses have you had that pertain specifically to music theory?

1% (1) None
2% (2) One
7% (3) Two
10% (4) Three
79% (5) Four or more

H. How many semester (quarter) courses have you had that pertain specifically to orchestration and instrumentation?

40% (1) None
35% (2) One
16% (3) Two
4% (4) Three
4% (5) Four or more

I. How many semester (quarter) courses have you had that pertain specifically to music history and literature?

1% (1) None
8% (2) One
23% (3) Two
21% (4) Three
47% (5) Four or more

ADVANCED PHILOSOPHY TEST

Content

The members of the committee responsible for the test recognize that the subject matter of philosophy is itself a topic of philosophical debate and that the subject can be taught in many different ways. Therefore, the test questions cover a wide range of material and approaches to philosophy so as to reflect the diversity of curriculums and student preparation.

A principal aim of the examination is to test the student's understanding. To this end questions have been devised that will favor the student who has read both widely and critically in philosophy and questions avoided that can be answered by a student who has chosen to rely on outlines and summaries in preparing for the examination. Thus, although some of the questions can be answered by drawing on purely factual information, the emphasis is on analysis, interpretation, and reasoning.

Most of the test is devoted to figures and problems in Western philosophy, with only an occasional question on Oriental systems of thought. About one-third of the questions assess philosophic reasoning and application of logical principles. The greatest historical emphasis is given to the period between 1600 and 1900, but the student is expected to show some knowledge of the ancient, medieval, and contemporary periods as well. The medieval period receives relatively little emphasis.

Various kinds of questions are used to test the student's competence in the broad areas of ethics, social and political philosophy, logic and philosophy of language, metaphysics and philosophy of mind, and epistemology. The fields of philosophy of science, aesthetics, philosophy of religion, and philosophy of history receive lesser emphasis. Some questions require students to demonstrate their ability to grasp the implications of ideas; others require them to identify important ideas of particular philosophers, solve problems in logic, or choose appropriate definitions of philosophical terms. Another type of question requires students to read a passage, the source of which may or may not be identified, and to choose an answer on the basis of their interpretation of the evidence supplied in the text or by applying what they have learned about the author of the passage.

There has been little attempt to include current moral and social problems as subject matter for philosophic discussion. This omission should not be construed as a judgment on the validity of approaching undergraduate philosophy through considering contemporary issues. Rather, it reflects the difficulties of anticipating changes in current concerns and of formulating questions in such a way that the students' own views would not impede their reading or answering correctly.

The establishment of subscores for factual information and philosophic reasoning is currently being planned.

Responses to Background Questions, 1970–71 (N = 1,570)

A. At what point are you in your studies?

3% (1) I am in or have just completed my junior year of undergraduate study.
65% (2) I am an undergraduate senior.
18% (3) I have a bachelor's degree but am not presently enrolled in graduate school.
8% (4) I am in or have just completed my first year of graduate study.
5% (5) I am in or have completed my second year of graduate study.

B. What graduate degree do you intend to seek?

3% (1) I do not plan to pursue graduate study.
3% (2) I plan to pursue graduate work but not to obtain a graduate degree.
15% (3) I plan to obtain terminal M.A., M.S., or other degree at the master's level.
28% (4) I plan to obtain M.A., M.S., or other master's level degree leading to a doctoral degree.
49% (5) I plan to obtain Ph.D., Ed.D., or other degree at the doctoral level.

C. Which of the following most accurately describes your preparation in philosophy?

22% (1) Philosophy major with honors or independent or advanced study
59% (2) Philosophy major
8% (3) Course work equivalent to that of a philosophy major
9% (4) Philosophy minor
2% (5) One course or none

D. If you are *not* a philosophy major, what is (was) your undergraduate major?

4% (1) A natural science or mathematics
3% (2) A social science (including psychology)
4% (3) A language or a literature or one of the arts
4% (4) History, politics, or government (including interdisciplinary area studies)
5% (5) Other

E. Which of the following best describes the material of the introductory course in philosophy that you took?

41% (1) Typical problems from different branches and periods of philosophy
23% (2) The doctrines of a very few major philosophers
8% (3) One period in the history of philosophy
11% (4) One of the areas of philosophy
12% (5) A systematic exposition of philosophical positions or schools

F. How many semester (quarter) courses have you had in logic?

17% (1) None
27% (2) One only, in traditional (nonsymbolic) logic
26% (3) One only, in symbolic logic
9% (4) One only, covering traditional logic and some scientific method
19% (5) At least two, including both traditional and symbolic logic

G. How many semester (quarter) courses have you had in the history of philosophy?

9% (1) None
10% (2) One only, in ancient philosophy
6% (3) One only, in modern philosophy
37% (4) A full year, covering ancient, medieval and modern philosophy
35% (5) A full year or more, including contemporary philosophy

H. Which of the following best describes your exposure to Oriental systems of thought?

42% (1) No course work and no independent reading
29% (2) Independent reading only
12% (3) Courses offering philosophy credit plus independent reading
3% (4) Courses not offering philosophy credit plus independent reading
13% (5) Course(s) in comparative religion or world religions

I. In which of the following areas are you LEAST prepared to answer questions?

16% (1) Social and political philosophy
34% (2) Aesthetics
35% (3) Philosophy of science
3% (4) Metaphysics
10% (5) Philosophy of religion

Philosophy Validity Data

Lannholm, Marco, and Schrader (1968) reported on a group composed of 42 students first enrolled in a particular graduate department of philosophy between the fall of 1957 and June 1960. The predictors were scores on the GRE Advanced Philosophy Test and the GRE Aptitude Test. The criterion was success in graduate study, defined as having earned the Ph.D. or being still enrolled and rated by faculty members as outstanding or superior in the fall of 1963. The results are reported in Table 17.

Table 17: Validities for 42 Philosophy Students

Predictors	r-biserial Correlation with Criterion	Mean Performance on Predictor	Standard Deviation of Performance on Predictor
GRE Advanced Philosophy	.11	779	101
GRE Verbal Ability	.17	708	73
GRE Quantitative Ability	.44	646	92

ADVANCED PHYSICS TEST

Content

The purpose of the test is to assess the students' understanding of fundamental principles and their ability to apply these principles in the solution of problems.

The approximate percentages of questions on content topics are as follows:

TOPIC	PERCENTAGE OF QUESTIONS
1. Classical mechanics, including Lagrangian and Hamiltonian formulation	18
2. Fundamentals of electromagnetism, including Maxwell's equations	18
3. Atomic physics	15
4. Physical optics and wave phenomena	10
5. Quantum mechanics	10
6. Special relativity	7
7. Thermodynamics and statistical mechanics	7
8. Laboratory methods	5
9. Nuclear and particle physics	4
10. Solid state physics	4
11. Miscellaneous	2

Responses to Background Questions, 1970–71 (N = 3,907)

A. At what point are you in your studies?

3% (1) I am in or have just completed my junior year of undergraduate study.
67% (2) I am an undergraduate senior.
12% (3) I have a bachelor's degree but am not presently enrolled in graduate school.
9% (4) I am in or have just completed my first year of graduate study.
7% (5) I am in or have completed my second year of graduate study.

B. What graduate degree do you intend to seek?

2% (1) I do not plan to pursue graduate study.
1% (2) I plan to pursue graduate work but not to obtain a graduate degree.
14% (3) I plan to obtain terminal M.A., M.S., or other degree at the master's level.
25% (4) I plan to obtain M.A., M.S., or other master's level degree leading to a doctoral degree.
57% (5) I plan to obtain Ph.D., Ed.D., or other degree at the doctoral level.

C. In which of the following fields did you major as an undergraduate?

85% (1) Physics
4% (2) Mathematics
5% (3) Engineering
1% (4) Chemistry
4% (5) Other

D. With respect to graduate schools, what is your reason for taking this test?

16% (1) To gain admission to a graduate school only
8% (2) To secure financial assistance from a graduate school only
70% (3) To gain admission to graduate school and to secure financial assistance

E. Are you taking this test in order to secure financial assistance from the National Science Foundation?

26% (1) Yes
69% (2) No

Validity Data

1. The group involved in a study by Michels (1966) was composed of 72 students who entered graduate institution A and 52 who entered institution B in the fall of 1962. The predictors were scores on the GRE Advanced Physics Test and the GRE Aptitude Test (verbal and quantitative). The criteria were a faculty ranking index, RI_1, of the students' performance based largely on grades in the first year of graduate study and a faculty ranking index, RI_3, based on performance in the first three years of graduate school. For institution A, which admitted 24 students with Advanced Physics Test scores below 600, the relations between predictors and criteria were strong, but for institution B, which admitted practically no students with Advanced Physics Test scores below 600, the relations between predictors and criteria were relatively weak. The probabilities of finding the relations observed in the absence of any actual correlation between the predictors and criteria are shown in Table 18.

Table 18: Probabilities of Finding Relations Observed between Predictors and Criteria in Physics for Institution A and B if the Correlations between Predictors and Criteria were Zero

		Probability	
Predictor	Criterion	Institution A	Institution B
GRE Advanced Physics	RI_1	.023	.61
GRE Verbal Ability	RI_1	.31	.50
GRE Quantitative Ability	RI_1	.024	.62
Sum of Advanced Physics and Quantitative Ability Scores	RI_1	$<.001$	.82
GRE Advanced Physics	RI_3	$<.001$	.82

It can be concluded that the Advanced Physics and quantitative ability scores are useful in distinguishing between outstanding and poor physics graduate students, but not very useful for distinguishing among various levels of outstanding students.

2. The group included in a study by Voorhees (1960) was composed of 68 graduate students admitted to the Department of Physics of the University of Chicago from the fall of 1950 through the fall of 1956. The predictor was the GRE Advanced Physics Test score and the criterion was success in graduate study. Successful students were those who obtained the Ph.D. or had passed the

Ph.D. candidacy examination. Approximately 92 percent of those with scores of 600 or above on the Advanced Physics Test were in the successful group. Only 53 percent of those with scores below 600 were successful.

3. The group included in a study by Lannholm, Marco, and Schrader (1968) was composed of 39 students first enrolled in a particular graduate department of physics between the fall of 1957 and June 1960. The predictors were scores on the GRE Advanced Physics Test and the GRE Aptitude Test. The criterion was success in graduate study, defined as having earned the Ph.D. or being still enrolled and rated by faculty members as outstanding or superior in the fall of 1963. The results are reported in Table 19. Some of the reasons for unexpectedly low or negative correlations between tests and performance are explained in Chapter 6.

Table 19: Validities for 39 Physics Students

Predictor	r-biserial Correlation with Criterion	Mean Performance on Predictor	Standard Deviation of Performance on Predictor
GRE Advanced Physics	.03	603	107
GRE Verbal Ability	.27	633	96
GRE Quantitative Ability	.01	704	81

4. The group in another study by Lannholm et al. (1968) was composed of 38 students first enrolled in a particular graduate department of physics between the fall of 1957 and June 1960. The predictors were scores on the GRE Advanced Physics Test and undergraduate grade-point average. The criterion was success in graduate study, defined as having earned the Ph.D. or being still enrolled and rated by faculty members as outstanding or superior in the fall of 1963. The results are reported in Table 20.

Table 20: Validities for 38 Physics Students

Predictor	r-biserial Correlation with Criterion	Mean Performance on Predictor	Standard Deviation of Performance on Predictor
GRE Advanced Physics	.68	621	121
UGPA	.40	3.13	.45

5. Roberts (1970) studied the records of 27 students who had enrolled at Wake Forest University from June 1964 to June 1970 for graduate study in physics and who had completed at least nine hours of graduate work. The correlations between graduate grade-point averages and GRE test scores were −.50 for verbal ability, −.01 for quantitative ability, and .30 for Advanced Physics.

6. The subjects for a study by Creager (1965) were 600 male applicants for National Science Foundation fellowships in 1955 and 1956. The predictors were scores on the GRE Aptitude Test (verbal and quantitative) and the Advanced Physics Test. One criterion was time lapse between attainment of the B.A. and the Ph.D., coded as shown below:

B.A.-Ph.D. Time Lapse (in years):	Less than 4	4	5	6	7	8	9	No Ph.D. by Aug. '64
Coded Variable:	1	2	3	4	5	6	7	8

A second criterion was the dichotomous variable of attaining or not attaining a Ph.D. by August 1964. The third criterion was the dichotomous variable of attaining or not attaining a Ph.D. in the average time taken to attain a Ph.D. in the field. The relationships between predictors and criteria are shown in Table 21.

Table 21: Validities of GRE against Doctorate Attainment for 600 Males Who Were Applicants for National Science Foundation Fellowships in Physics in 1955 and 1956

Predictors	Criteria				
	Reflected B.A.-Ph.D. Time Lapse[1]	Ph.D. by 1964		Ph.D. in Average Time	
		Point Biserial	Biserial	Point Biserial	Biserial
GRE Verbal Ability	.16	.15	.19	.12	.16
GRE Quantitative Ability	.26	.26	.33	.23	.30
GRE Advanced Physics	.34	.32	.31	.30	.39
Composite	.35	.33	.43	.32	.42

[1] Correlations between the coded variable for B.A.-Ph.D. time lapse given above and the predictors with the signs reversed.

ADVANCED POLITICAL SCIENCE TEST

Content

In preparing the test, diversity of curriculums and backgrounds of students are taken into account. The questions are drawn from the courses of study most commonly offered.

The distribution of questions among subfields of the discipline, and within each subfield, according to skills, processes, and approaches is reviewed yearly by the committee of examiners. To facilitate the process of allocating questions, the committee has developed the following test specifications. The specifications are an approximation of the content breakdown, but the cell percentages do not serve as rigid guidelines for the selection of questions.

Specifications for the GRE Advanced Political Science Test

	International Relations	Comparative Political Systems	American Government	Generic	TOTAL
Law	2.5%	2.5%	2.5%	2.5%	10.0%
Politics and Political Behavior Legislative Executive and Administrative Judicial Elections and Voting Attitudes	7.5%	12.5%	17.5%	5.0%	42.5%
Government Structure; Organization	2.5%	5.0%	5.0%		12.5%
Theory and Approaches	1.5%	1.5%	2.0%	5.0%	10.0%
Methodology	1.0%	1.0%	1.0%	12.0%	15.0%
TOTAL	15.0%	22.5%	28.[illegible]	24.5%	90.0%
History of Political Thought		10.0%			10.0%
TOTAL					100.0%

Although specific information is needed to answer many of the questions, most questions are not limited solely to recall.

Responses to Background Questions, 1970–71 (N = 5,314)

A. At what point are you in your studies?

3% (1) I am in or have just completed my junior year of undergraduate study.
62% (2) I am an undergraduate senior.
21% (3) I have a bachelor's degree but am not presently enrolled in graduate school.
8% (4) I am in or have just completed my first year of graduate study.
5% (5) I am in or have completed my second year of graduate study.

B. What graduate degree do you intend to seek?

3% (1) I do not plan to pursue graduate study.
2% (2) I plan to pursue graduate work but not to obtain a graduate degree.
39% (3) I plan to obtain terminal M.A., M.S., or other degree at the master's level.
30% (4) I plan to obtain M.A., M.S., or other master's level degree leading to a doctoral degree.
25% (5) I plan to obtain Ph.D., Ed.D., or other degree at the doctoral level.

C. What is (was) your undergraduate major?

80% (1) Political science
7% (2) Other social science
1% (3) Mathematics or a natural science
6% (4) History or other humanities
6% (5) Other

D. What is (was) your undergraduate minor?

7% (1) Political science
19% (2) Other social science
2% (3) Mathematics or a natural science
28% (4) History or other humanities
41% (5) Other or not applicable

E. How would you classify the distribution of your course work in political science?

14% (1) Highly concentrated in one field
26% (2) Directed toward one field
20% (3) Fairly evenly divided between two major fields
13% (4) Distributed among three fields
25% (5) Distributed equally across the discipline

F. In which of the following fields did you concentrate your undergraduate work in political science?

34% (1) American government and politics (including public law)
5% (2) Urban affairs
15% (3) Comparative government and politics
26% (4) International relations
12% (5) Political theory (normative)

G. If you were given the opportunity to select one of the following tests or methods of reporting scores for graduate school admission, which would you prefer?

9% (1) The current Advanced Political Science Test covering all fields, with a single score reported
19% (2) The current examination, with separate scores reported for each major field convered on the test
21% (3) A shortened version of the present test plus your choice of one of three or four optional field tests, with two separate scores reported
36% (4) Your choice of two of the following field tests with two scores reported: American Government and Politics, Comparative Government and Politics, International Relations, Normative Political Theory
12% (5) Other

H. Have you taken a political science methodology and/or statistics course?

61% (1) I have taken neither a methodology nor a statistics course.
16% (2) I have taken a methodology course only.
10% (3) I have taken a statistics course only.
5% (4) I have taken one course combining methodology and statistics.
5% (5) I have taken both a methodology and a statistics course.

I. Have you ever done independent research requiring the collection, processing, and interpretation of data?

39% (1) Yes
59% (2) No

ADVANCED PSYCHOLOGY TEST

Content

The questions in the test are drawn from courses of study most commonly offered within the broadly defined field of psychology. Questions in the test often require the student to identify psychologists associated with particular theories or conclusions and to recall information from psychology courses. In addition, some questions require analyzing relationships, applying principles, drawing conclusions from experimental data, and evaluating experiments.

Although the test offers only two subscores, there are questions in three content categories, as follows:

1. Experimental or natural science oriented, with questions distributed about equally among learning, physiological and comparative, and perception and sensory psychology.
2. Social or social science oriented, with questions distributed about equally among personality, clinical and abnormal, developmental, and social psychology.
3. General, including historical and applied psychology, measurement, and statistics.

A separate subscore is reported *for only the first two of these categories*. Evidence from students' performance on test questions shows that questions within each of the two subscore categories are more closely related to each other than are questions in different categories. Each of the subscores reported for the test is based on approximately 40 percent of the questions in the entire test.

Research Related to Additional Subscores on the Advanced Psychology Test

Two subscores, Experimental Psychology and Social Psychology, are currently reported for the Advanced Psychology Test. The GRE Board has considered, however, that "it is both desirable and feasible to report more detailed and useful part-score information on the basis of the Advanced Tests." The recommendation that the reporting of additional subscores on the Advanced Tests be investigated stemmed from a feeling on the part of both the Board and several of the committees of examiners that more than a single score should be produced from three hours of testing and that tests would be much more useful—particularly to students—if they could indicate strengths and weaknesses in the several subfields of each content area. Also, the Board and the committees recognized that such subscores could be valuable for many counseling and placement decisions.

However, in spite of the widespread agreement about the desirability of reporting as many subscores as possible, there are a number of areas of concern—including the reliability and independence of the subscores—that center on the eventual use of the subscores. If the use of subscores were restricted to placement and guidance, the importance of these concerns would diminish. Placement and counseling decisions are reversible, whereas admission decisions generally are not; therefore, much lower standards of statistical adequacy would be acceptable if subscores were not used for admission decisions. This would enable many more subscores to be reported, while allowing test committees to give continued emphasis to the various elements in their disciplines.

A study was designed to investigate the number of logically meaningful subscores that could be generated from the Advanced Psychology Test if the statistical standards for subscores used for admission purposes are relaxed. The study examined the reliability and independence of subscores based on the eight major content areas of the Advanced Psychology Test as an initial step in determining the extent to which the propositions endorsed by the GRE Board were conceptually and psychometrically feasible for extension to the design and administration of the GRE Advanced Tests.

On the assumption that a subscore might be developed in each of the major content areas established by the committee as part of its test specifications, content analysis was used to define the structure of two forms of the GRE Advanced Psychology Test. These areas, each of which would provide valuable data for counseling graduate students, include: (1) Personality, (2) Learning, (3) Measurement, (4) Developmental Psychology, (5) Social Psychology, (6) Physiological and Comparative Psychology, (7) Perception and Sensory Psychology and (8) Clinical and Abnormal Psychology.

The results of the study show that each of the eight subscores appears to meet the criteria of independence set for subscores in the GRE Program. This confirms the validity of the committee of examiners' belief that subscores based on the eight content areas would be about as independent as the two subscores being reported at present. The committee feels that subscores based on the content areas would be by far the most useful ones for purposes of guidance and placement because the curriculum tends to be organized in the same way.

In addition, it appears that subscores based on the content areas defined by the committee have considerable potential to provide information about students with unusually high or low scores for use in guidance and placement.

Factor analysis was used as a second and less subjective method of examining the structure of the two forms of the GRE Advanced Psychology Test, and it offers an independent description of the observed data.

Factor analytic techniques were used to investigate two separate questions:

1. To what extent do the items in each content-determined subscore appear to be measuring a single ge[illegible] factor?
2. Do other logically valid and potentially useful groupings (subscores) of the items exist in the data?

The committee felt that the results of the factor analysis of the two test forms were basically consistent in identifying factors that measure primarily knowledge of facts, knowledge of theories, and powers of interpretation and analysis. However, the committee did not believe that subscores based on the factor analysis would be useful for guidance and placement. The point was made that the curriculum is organized along content-determined lines, and that students need to know their strengths and weaknesses in those terms. The committee felt that the results of the factor analysis in no way limited the usefulness of subscores based on content areas.

The committee's content analysis is a logical way of relating test content to the curriculum, and is intended to ensure that all the major areas of the curriculum are appropriately represented in the test. In view of the relative independence of the subscores based on content areas, the fact that each content area did not emerge as a separate factor in the factor analysis was not disturbing. Such a result is consistent with the fact that content areas are not learned independently, that many require introductory courses covering all the major areas of psychology, and that much psychological theory is applicable across content areas.

The future of such subscore reporting depends, however, on findings for other Advanced Tests and significant changes in the GRE Program to report such subscores in a way that would prevent their use for other than counseling purposes.

Responses to Background Questions, 1970–71
(N = 17,578)

A. At what point are you in your studies?

4% (1) I am in or have just completed my junior year of undergraduate study.
65% (2) I am an undergraduate senior.
18% (3) I have a bachelor's degree but am not presently enrolled in graduate school.
7% (4) I am in or have just completed my first year of graduate study.
5% (5) I am in or have completed my second year of graduate study.

B. What graduate degree do you intend to seek?

2% (1) I do not plan to pursue graduate study.
1% (2) I plan to pursue graduate work but not to obtain a graduate degree.
23% (3) I plan to obtain terminal M.A., M.S., or other degree at the master's level.
27% (4) I plan to obtain M.A., M.S., or other master's level degree leading to a doctoral degree.
45% (5) I plan to obtain Ph.D., Ed.D., or other degree at the doctoral level.

C. If you are now a college senior, which of the following best describes your educational experience and your plans with respect to the graduate study of psychology? (If you are not a senior, mark 5.)

50% (1) I am an undergraduate major in psychology and I plan to do graduate work in psychology.
10% (2) I am an undergraduate major in psychology and I plan to do graduate work in some other field related to psychology.
2% (3) I am an undergraduate major in psychology and I plan to do graduate work in some field not related to psychology.
4% (4) I am not an undergraduate major in psychology but I plan to do graduate work in psychology.
27% (5) Other

D. In what general area would you classify your undergraduate major?

79% (1) Social science
5% (2) Biological science
2% (3) Physical science
1% (4) Mathematics
12% (5) Other

E. What is (was) your undergraduate major?

83% (1) Psychology
1% (2) Philosophy
2% (3) Sociology
2% (4) Education
11% (5) Other

F. In what area of psychology have you had the most course work?

31% (1) Clinical or abnormal
8% (2) Educational
29% (3) Experimental
13% (4) Social
18% (5) Other

G. Which of the following best describes your work in these three courses: general (or introductory) psychology, experimental psychology, and statistics?

14% (1) General psychology only
1% (2) Experimental psychology only
10% (3) General psychology and experimental psychology only
64% (4) General psychology, experimental psychology, and statistics
10% (5) Other

H. How recently have you had a college or graduate course in psychology?

77% (1) During the current academic year
13% (2) During the previous academic year
5% (3) Two or three years ago
2% (4) Four or five years ago
2% (5) Other

I. In what area of psychology, if any, do you plan to pursue your career?

44% (1) Clinical or abnormal
13% (2) Educational
11% (3) Experimental
10% (4) Social
20% (5) Other, or not in psychology

Validity Data

1. Rock (1972) reported on a group composed of 778 applicants for National Science Foundation fellowships. Most applied for NSF fellowships in 1958–61. The predictors were scores on the GRE Advanced Psychology Test and the GRE Aptitude Test (verbal and quantitative), undergraduate grade-point average and an average rating of reference letters. The criterion was attainment of the doctorate by June 1968. The group was split into two random halves; the validity coefficients for each half are shown in Table 22.

2. A group reported on by Lorge (1960) was composed of 165 graduate students majoring in Psychological Foundations of Education at Teachers College, Columbia University. Using the score obtained on the doctoral written examination as the criterion, correlations of .41 with the GRE Advanced Psychology Test score, .63 with the GRE verbal ability score, and .32 with the GRE quantitative ability score were found.

3. The group involved in a study by Sistrunk (1961) was composed of 73 graduate students in psychology at the University of Miami. The correlations between the predictor (GRE Advanced Psychology Test score) and the criterion (department examination) was .56.

4. The group included in a study by Lannholm, Marco, and Schrader (1968) was composed of 47 students first enrolled in a particular graduate department of psychology between the fall of 1957 and June 1960. The predictors were scores on the GRE Advanced Psychology Test and the GRE Aptitude Test. The criterion was success in graduate study, defined as having earned the Ph.D. or being still enrolled and rated by faculty members as outstanding or superior in the fall of 1963. The results are shown in Table 23.

Table 22: Validities Using the Criterion of Attainment of the Doctorate for 778 National Science Foundation Fellowship Applicants in Psychology in 1958–61, Split into Two Random Halves

	n = 380			n = 398		
Predictors	r-biserial Correlation with Criterion	Predictor Performance		r-biserial Correlation with Criterion	Predictor Performance	
		Mean	Standard Deviation		Mean	Standard Deviation
GRE Advanced Psychology[1]	.19	60.98	8.90	.24	60.87	9.05
GRE Verbal Ability[1]	.12	63.52	8.25	.19	63.47	9.29
GRE Quantitative Ability[1]	.33	59.89	11.34	.14	60.96	10.82
UGPA[2]	.02	241.70[3]	40.10[3]	.02	236.78	42.96
Reference Letters[4]	.16	43.66	8.36	.14	43.93	8.49

[1] Scaled score with third digit dropped
[2] On a four-point scale multiplied by 100
[3] n = 462
[4] Zero to 6 multiplied by 10

Table 23: Validities for 47 Psychology Students

Predictor	r-biserial Correlation with Criterion	Mean Performance on Predictor	Standard Deviation of Performance on Predictor
GRE Advanced Psychology	.35	685	49
GRE Verbal Ability	.28	705	62
GRE Quantitative Ability	.27	678	81

5. The group involved in another study by Lannholm et al. (1968) was composed of 38 students first enrolled in a particular graduate department of psychology between the fall of 1957 and June 1960. The predictors were scores on the GRE Advanced Psychology Test and the GRE Aptitude Test. The criterion was success in graduate study, defined as having earned the Ph.D. or being still enrolled and rated by faculty members as outstanding or superior in the fall of 1963. The results are shown in Table 24.

Table 24: Validities for 38 Psychology Students

Predictor	r-biserial Correlation with Criterion	Mean Performance on Predictor	Standard Deviation of Performance on Predictor
GRE Advanced Psychology	–.11	640	65
GRE Verbal Ability	–.02	685	65
GRE Quantitative Ability	.18	608	99

6. A third study by Lannholm et al. (1968) involved a group composed of 26 students first enrolled in a particular graduate department of psychology between the fall of 1957 and June 1960. The predictors were scores on the GRE Advanced Psychology Test and the GRE Aptitute Test and undergraduate grade-point average. The criterion was success in graduate study, defined as having earned the Ph.D. or being still enrolled and rated by faculty members as outstanding or superior in the fall of 1963. The results are shown in Table 25.

Table 25: Validities for 26 Psychology Students

Predictor	r-biserial Correlation with Criterion	Mean Performance on Predictor	Standard Deviation of Performance on Predictor
GRE Advanced Psychology	.29	607	75
GRE Verbal Ability	.27	603	90
GRE Quantitative Ability	.45	548	96
UGPA	.11	2.96	.44

7. A fourth group studied by Lannholm et al. (1968) was composed of 26 students first enrolled in a particular graduate department of psychology between the fall of 1957 and June 1960. The predictors were scores on the GRE Advanced Psychology Test and the GRE Aptitude Test. The criterion was success in graduate study, defined as having earned the Ph.D. or being still enrolled and rated by faculty members as outstanding or superior in the fall of 1963. The results are shown in Table 26.

Table 26: Validities for 26 Psychology Students

Predictor	r-biserial Correlation with Criterion	Mean Performance on Predictor	Standard Deviation of Performance on Predictor
GRE Advanced Psychology	–.35	621	66
GRE Verbal Ability	–.27	653	70
GRE Quantitative Ability	–.24	597	79

Table 27: Validities for 31 Students in Psychology at New York University

Predictors	Criteria		Performance on Predictors	
	Ph.D. Attainment[2]	Percentage of A's	Mean	Standard Deviation
GRE Advanced Psychology[1]	.66	.44	.88	.62
GRE Verbal Ability[1]	.17	–.01	1.56	.72
GRE Quantitative Ability[1]	.22	.17	1.07	.84
MAT	–.07	–.28	.39	.97
Overall UGPA	.10	.11	2.93	.41
Psychology UGPA	.04	.28	3.34	.49
Number of Undergraduate Psychology Courses	.32	.02	7.33	3.79
Performance on Criteria				
Mean	9.52	.51		
Standard Deviation	38.16	33.64		

[1] GRE Scores are neither raw nor scaled scores.

[2] Receiving Ph.D. = 1; not receiving Ph.D. = 0.

8. The subjects of a study by Ewen (1969) were 31 men enrolled in psychology at New York University no earlier than the fall of 1960. The predictors were scores on the GRE Advanced Psychology Test, the GRE Aptitude Test, and, the Miller Analogies Test, overall undergraduate grade-point average, psychology undergraduate grade-point average, and number of undergraduate psychology courses taken. The criteria were attainment of the Ph.D. and percentage of "A" grades in graduate school. Of the 31 subjects, 16 earned the Ph.D. and 15 were dropped or withdrew. The zero-order correlations between predictors and criteria are shown in Table 27.

9. The subjects of a study reported on by Hackman, Wiggins, and Bass (1970) were 42 students who began doctoral work in psychology at the University of Illinois in 1963. The predictors included scores on the GRE Advanced Psychology Test and the GRE Aptitude Test, the number of languages spoken, the number of languges read, the number of semester hours of language taken as an undergraduate, undergraduate grade-point leverage in all courses in the junior and senior years, and the quality of the undergraduate institution as judged by a faculty committee. The criteria at the end of the first year of graduate study included grades, two student assessments of their own progress, and two faculty judgments of student progress. The first student assessment dealt with how rapidly they thought they were progressing toward a doctorate, and the second dealt with whether or not they planned to continue graduate study at Illinois. The first faculty rating was made by teachers who had had the students in a course, the second was made by heads of departmental divisions. Six years after beginning graduate work all students had either earned a Ph.D. or had withdrawn. All students were rated by faculty members on a 9-point scale of success to provide a long-term criterion. The results are shown in Table 28.

Table 28: Correlations Between Predictors and Criteria for 42 Students in Psychology at the University of Illinois

Predictors	Criteria					
	End of First Year					After 6 Years
	Student Self Assessment			Faculty Ratings		
	Grades	Speed to Degree	Plans to Continue	Teachers	Dept. Heads	Success Rating
GRE Advanced Psychology	.23	.23	.16	.08	.12	–.11
GRE Verbal Ability	.22	.45[1]	.23	.21	.20	.19
GRE Quantitative Ability	.15	.40[1]	.03	.29	.23	.32[1]
No. Languages Spoken	.04	–.04	–.39[1]	–.04	.10	–.21
No. Languages Read	.19	.07	–.47[1]	.01	.28	–.25
Hours Language Study	–.06	–.20	–.25	–.03	–.02	–.34[1]
UGPA Last 2 Years	.28	.02	–.04	–.08	.05	–.22
Quality of Undergraduate Institution	.15	.30[1]	.16	.31[1]	.08	.43[1]

[1] Significant at the .05 level.

10. A group reported on by Newman (1968) was composed of 66 graduate students studying for advanced degrees in the Department of Psychology at Washington State University. The predictors were scores on the GRE Advanced Psychology Test and the GRE Aptitude Test for 27 students. For the remaining 39 students the predictors were only the two Aptitude Test scores. The criterion was graduate grade-point average. The results are shown in Table 29.

Table 29: Validities for 66 Psychology Students at Washington State University

Predictor	Correlation with Criterion	Mean Performance on Predictor	Standard Deviation of Performance on Predictor
GRE Advanced Psychology (n = 27)	.09	607	56
GRE Verbal Ability	.08	582	92
GRE Quantitative Ability	.21[1]	535	104

[1] Significant at the .05 level.

11. The subjects for a study by Creager (1965) were 99 applicants for National Science Foundation fellowships in 1955 and 1956. The predictors were scores on the GRE Aptitude Test (verbal and quantitative) and the Advanced Psychology Test. One criterion was time lapse between attainment of the B.A. and the Ph.D., coded as shown below.

B.A.-Ph.D. Time Lapse (in years):	Less than 4	4	5	6	7	8	9	No Ph.D. by Aug. '64
Coded Variable:	1	2	3	4	5	6	7	8

A second criterion was the dichotomous variable of attaining or not attaining a Ph.D. by August 1964. The third criterion was the dichotomous variable of attaining or not attaining a Ph.D. in the average time taken to attain a Ph.D. in the field. The relationships between predictors and criteria are shown in Table 30.

Table 30: Validities of GRE against Doctorate Attainment for 99 Applicants for National Science Foundation Fellowship in Psychology in 1955 and 1956

Predictors	Criteria				
	Reflected B.A.-Ph.D. Time Lapse[1]	Ph.D. by 1964		Ph.D. in Average Time	
		Point Biserial	Biserial	Point Biserial	Biserial
GRE Verbal Ability	.13	.13	.18	.17	.24
GRE Quantitative Ability	.17	.13	.18	.16	.22
GRE Advanced Psychology	.33	.25	.34	.30	.42
Composite	.34	.25	.34	.30	.42

[1] Correlations between the coded variable for B.A.-Ph.D. time lapse given above and the predictors with the signs reversed.

12. The group included in a study by Mehrabian (1969) was composed of 79 students enrolled in the graduate psychology program at the University of California, Los Angeles. The predictors were scores on the GRE Advanced Psychology Test, the GRE Aptitude Test, and the Miller Analogies Test, overall undergraduate grade-point average, UGPA in the last two undergraduate years, number of undergraduate courses in mathematics and logic taken, faculty ratings of promise as a student and researcher, faculty rating of research orientation, admissions committee evaluation of promise, and admissions committee index of acceptability.

The criteria were faculty rating of graduate achievement, average grade in all first-year content courses, and average grade in all first-year statistics courses. The results are shown in Table 31.

Table 31: Correlations Between Predictors and Criteria for 79 Psychology Students at the University of California, Los Angeles

Predictors	Criteria		
	Faculty Rating Graduate Achievement	Average Grade in Content	Average Grade in Statistics
GRE Advanced Psychology	.48[1]	.53[1]	.61[1]
GRE Verbal Ability	−.14	.17	.12
GRE Quantitative Ability	.15	.27[1]	.48[1]
MAT	.19	.34[1]	.33[1]
Overall UGPA	.08	.10	.08
UGPA Last 2 Years	.13	.21	.14
Number of Mathematics and Logic Courses	.15	.11	.33[1]
Faculty Rating of Promise	.25[1]	.25[1]	.38[1]
Faculty Rating of Research Orientation	.19	.31[1]	.06
Admissions Committee Evaluation of Promise	.10	.22	.33[1]
Admissions Committee Index of Acceptability	.30[1]	.21	.20

[1] Correlation coefficients above .22 are significant at the .05 level.

ADVANCED SOCIOLOGY TEST

Content

The questions in the test are drawn from the courses of study most commonly offered in college curriculums. A few examples of course titles are theory, collective behavior, social institutions, introductory statistics, urban sociology, demography, human ecology, social structure and personality, criminology and juvenile delinquency, public opinion and propaganda, research methods, and the logic of sociological inquiry. The test aims at a balance among the many subfields of sociology, and the questions are distributed among the following areas:

AREA	PERCENT	AREA	PERCENT
Methodology and statistics	15	Theory	6
Social psychology	9	Stratification	5
Race and ethnic relations	8	Comparative sociology	4
Social change	8	Occupations and professions	4
Complex organization	6	Political sociology	4
Demography	6	Social organization	4
Deviance and social control	6	Urban/rural sociology	4
Marriage and the family	6	Collective behavior	2
		Human ecology	2
		Religion	1

The coverage of methodology and statistics in the test is actually greater than 15 percent, because a number of questions in the other areas enumerated above also require methodological or statistical skills.

Recall of specific information is required to answer many of the questions. The test, however, does not merely measure factual knowledge as such but instead draws upon such knowledge to test for ability to interpret the types of data typically encountered by sociologists and for an understanding of relationships.

Responses to Background Questions, 1970–71
(N = 1,739)

A. At what point are you in your studies?

3% (1) I am in or have just completed my junior year of undergraduate study.
67% (2) I am an undergraduate senior.
18% (3) I have a bachelor's degree but am not presently enrolled in graduate school
6% (4) I am in or have just completed my first year of graduate study.
4% (5) I am in or have completed my second year of graduate study.

B. What graduate degree do you intend to seek?

5% (1) I do not plan to pursue graduate study.
3% (2) I plan to pursue graduate work but not to obtain a graduate degree.
46% (3) I plan to obtain terminal M.A., M.S., or other degree at the master's level.
26% (4) I plan to obtain M.A., M.S., or other master's level degree leading to a doctoral degree.
18% (5) I plan to obtain Ph.D., Ed.D., or other degree at the doctoral level.

C. What was your undergraduate major field?

63% (1) Sociology only
17% (2) Sociology and another social science
5% (3) Sociology and the humanities
2% (4) Sociology and mathematics or a natural science
12% (5) Other

D. If your major was sociology, what was your undergraduate *minor* field?

38% (1) Another social science
5% (2) Education
4% (3) Science or mathematics
12% (4) Humanities
21% (5) Other

E. Have you had a course in contemporary sociological theory and/or a course in the history of social theory?

33% (1) No
12% (2) Yes, a course in contemporary theory
28% (3) Yes, a course in history of social theory
17% (4) Yes, a course combining contemporary theory and the history of social theory
9% (5) Yes, a course in contemporary theory and a course in the history of social theory

F. What is the highest level mathematics course you have taken in college?

36% (1) No mathematics course in college
36% (2) Algebra and trigonometry
16% (3) Elementary calculus
3% (4) Advanced calculus
1% (5) Courses beyond advanced calculus

G. Have you had a course in statistics?

64% (1) Yes
35% (2) No

H. Have you had a course in race and/or ethnic relations?

51% (1) Yes
48% (2) No

I. Have you had a course in demography or population?

22% (1) Yes
76% (2) No

Validity Data

Roberts (1970) studied the records of 24 students who had enrolled at Wake Forest University from June 1964 to June 1970 for graduate work in sociology and anthropology and who had completed at least nine hours of graduate work. The correlations between graduate grade-point averages and GRE scores were .16 for verbal ability, .17 for quantitative ability, and .69 for the Advanced Test. Presumably the Advanced Test was in sociology in most instances.

ADVANCED SPANISH TEST

Content

In determining the content of the Advanced Spanish Test, the committee of examiners must take into account (1) the diversity of subject matter and emphasis in undergraduate curriculums; (2) the areas of specialization most likely to be entered upon by graduate students; and (3) the abilities most relevant to the tasks graduate students are likely to encounter. Accordingly, the test contains questions in the following broad areas.

Language Proficiency and Knowledge. A certain number of questions focus directly on the student's mastery of correct structure and usage. A few questions in the field of descriptive and structural linguistics are also included. However, since the entire test is in Spanish, all questions basically involve the student's knowledge of the language.

The structure and usage questions will give some indication of a student's ability to write acceptable Spanish; oral skills, however, are not measured in the framework of this test. Proof of students' proficiency in speaking Spanish and in understanding spoken Spanish will, therefore, have to be obtained in other ways.

Literary History and Theory. Questions in this area test the student's familiarity with those Spanish and Spanish American authors whose works are most likely to be studied by undergraduate majors in Spanish. Although some questions are limited to factual recall, others probe deeper to gauge understanding of literary trends and ideas. Because of the diversity of undergraduate literature courses and reading programs, however, it is unlikely that any students will have read all the works represented in the test and, consequently, that they will be able to answer all questions in this area.

Other questions test familiarity with basic concepts and terms of literary theory.

Literary Interpretation and Insight. The ability to comprehend the meaning of literary works fully and to interpret them with sensitivity and insight is of particular importance to students preparing for an advanced degree in Spanish or Spanish American literature. A number of questions in the test give them an opportunity to demonstrate the skills they have acquired during their undergraduate studies and their aptitude in the area of literary interpretation. These questions deal with aspects of meaning or form in literary selections.

Culture and Civilization. The committee feels that some knowledge and understanding of Spanish and Spanish American culture and civilization is essential for the student entering upon graduate study. Accordingly, the test contains a certain number of questions touching on major aspects of history, geography, institutions, customs, ideas, and the arts in the Hispanic world.

In assigning relative weights to Peninsular and Spanish American subject matter, the committee bears in mind the growing attention given to Spanish America in undergraduate programs. Even so, to reflect the reality of undergraduate curriculums, somewhat more weight is allotted to Spain. As a general rule, structure and usage peculiar to any part of the Spanish-speaking world are avoided unless knowledge of them is to be explicitly tested.

The test yields a total score and the three subscores of 1) Interpretive Reading Skills; 2) Peninsular Topics; and 3) Spanish American Topics.

Responses to Background Questions, 1970–71
(N = 1,739)

A. At what point are you in your studies?

3% (1) I am in or have just completed my junior year of undergraduate study.
60% (2) I am an undergraduate senior.
20% (3) I have a bachelor's degree but am not presently enrolled in graduate school.
10% (4) I am in or have just completed my first year of graduate study.
6% (5) I am in or have completed my second year of graduate study.

B. What graduate degree do you intend to seek?

3% (1) I do not plan to pursue graduate study.
3% (2) I plan to pursue graduate work but not to obtain a graduate degree.
49% (3) I plan to obtain terminal M.A., M.S., or other degree at the master's level.
28% (4) I plan to obtain M.A., M.S., or other master's level degree leading to a doctoral degree.
15% (5) I plan to obtain Ph.D., Ed.D., or other degree at the doctoral level.

C. Was Spanish regularly spoken in your home when you were a child?

25% (1) Yes

74% (2) No

D. For what length of time have you studied in or lived in a Spanish-speaking country?

27% (1) Not at all

21% (2) Some, but less than three months

10% (3) Three to six months

13% (4) Six months to one year

28% (5) More than one year

E. What is (or was) your undergraduate major field?

81% (1) Spanish

3% (2) Another foreign language

14% (3) Other

F. If you majored in Spanish as an undergraduate, which of the following was most emphasized in your courses?

62% (1) Literature

14% (2) Language proficiency

5% (3) Civilization and culture (including area studies)

3% (4) Linguistics (history of language, structure of language)

5% (5) Other

G. In your undergraduate literature and/or civilization courses, what was the relative emphasis given to Spain and Spanish America?

58% (1) Greater emphasis was given to Spain.

12% (2) Greater emphasis was given to Spanish America.

25% (3) Spain and Spanish America received about equal emphasis.

References

Creager, J. A. *Predicting doctorate attainment with GRE and other variables* (Technical Report No. 25). Washington, D.C.: Office of Scientific Personnel, National Academy of Sciences—National Research Council, November 1965.

Eckhoff, C. M. Predicting graduate success at Winona State College. *Educational and Psychological Measurement*, 1966, *26*, 483–485.

Ewen, R. B. The GRE psychology test as an unobtrusive measure of motivation. *Journal of Applied Psychology*, 1969, *53*, 383–387.

Hackman, J. R., Wiggins, N., & Bass, A. R. Prediction of long-term success in doctoral work in psychology. *Educational and Psychological Measurement*, 1970, *30*, 365–374.

Johnson, H., & Thompson, E. *The Graduate Record Examinations at Sacramento State College* (Technical Bulletin No. 11, Student Personnel Division, Sacramento State College, 1962). Reported by G. V. Lannholm in GRE Special Report 68-1. Princeton, N.J.: Educational Testing Service, 1968.

Lannholm, G. V., Marco, G. L., & Schrader, W. B. *Cooperative studies of predicting graduate school success* (GRE Special Report 68-3). Princeton, N.J.: Educational Testing Service, 1968.

Lorge, I. *Relationship between Graduate Record Examinations and Teachers College, Columbia University, doctoral verbal examinations* (Letter to G. V. Lannholm dated September 21, 1960). Reported by G. V. Lannholm in GRE Special Report 68-1. Princeton, N.J.: Educational Testing Service, 1968.

Mehrabian, A. Undergraduate ability factors in relationship to graduate performance. *Educational and Psychological Measurement*, 1969, *29*, 409–419.

Michels, W. C. Graduate Record Examinations Advanced Physics Test as a predictor of performance. *American Journal of Physics*, 1966, *34* (9, Pt. 2).

Newman, R. I. GRE scores as predictors of GPA for psychology graduate students. *Educational and Psychological Measurement*, 1968, *28*, 433–436.

Office of Educational Research. *Study of GRE scores of geology students matriculating in the years 1952–1961* (RP-Abstract, Yale University, 1963). Reported by G. V. Lannholm in GRE Special Report 68-1. Princeton, N.J.: Educational Testing Service, 1968.

Roberts, P. T. *An analysis of the relationship between Graduate Record Examination scores and success in the Graduate School of Wake Forest University*. Unpublished master's thesis, Wake Forest University, 1970.

Rock, D. A. *The prediction of doctorate attainment in psychology, mathematics, and chemistry* (GRE Board Report 69-6). Princeton, N.J.: Educational Testing Service, 1972.

Roscoe, J. T., & Houston, S. R. The predictive validity of GRE scores for a doctoral program in education. *Educational and Psychological Measurement*, 1969, *29*, 507–509.

Sistrunk, F. *The GREs as predictors of graduate school success in psychology* (Letter to G. V. Lannholm dated October 3, 1961). Reported by G. V. Lannholm in GRE Special Report 68-1. Princeton, N.J.: Educational Testing Service, 1968.

Voorhees, H. R. *Relationship between scores on Graduate Record Examinations and graduate school performance in physics*. Unpublished manuscript cited by G. V. Lannholm in GRE Special Report 60-3. Princeton, N.J.: Educational Testing Service, 1960.

Williams, J. D., Harlow, S. D., & Grab, D. A longitudinal study examining prediction of doctoral success: Grade point average as criterion, or graduation vs. nongraduation as criterion. *Journal of Educational Research*, December 1970, *64*, 161–164.

INDEX

Advanced Tests 2, 24-31.
    See also specific subjects.
  assembly of, 6, 29
  committees of examiners, 10, 26-27, 34
  content specifications of, 27
  correlations among subscores, 30
  correlations with Aptitude Test scores, 30-31
  criteria for developing, 24
  development of, 5-6, 24, 31
  equating of, 34-36
  examinee volume in, 48
  format of, 25-26
  formula scoring, 4-5
  general characteristics of, 4-8
  item analysis of, 7, 39-41
  pretesting of, 7, 29
  purpose of, 4
  quality control of, 7
  reliability of, 28-29, 39
  rescaling study, 38-39
  review of, 5-6
  scaled-score system, 32-33
    stability of, 37-38
  speededness, 29
  statistical specifications and characteristics of, 28-31
  subscores, 29-30, 36, 43, 47
  test analysis of, 41-46
  tests available, list, 31
  uses of, 24-25
  validity of, 52-53, 55, 57.
    See also specific subjects.
Al-Amire, H., 61
Alexakos, C. E., 60
Alternate forms, 32
Altman, R. A., 51
Analogies, 10, 11
Analysis of explanations, 16-17
Analytical ability measure, 10, 16-18, 19, 22-23, 32, 33-34, 54, 55, 64-67
  analysis of explanations, 16-17
  analytical reasoning, 18
  content specifications of, 18
  development of, 10
  format of, 10
  logical diagrams, 17
  question types not selected for use, 64-67
  reliability of, 19
  scaling of, 33-34
  validity of, 22-23
Analytical reasoning, 18
Angoff, W. H., 33, 59, 60
Antonyms, 11
Aptitude Test, 9-23.
    See also Analytical ability measure; Quantitative ability measure; Verbal ability measure.
  assembly of, 6
  changes in, 9-10
  content characteristics of, 10
  correlations, internal, 18-19
  correlations with Advanced Test scores, 30-31
  descriptive statistics, 48-51
  development of, 5-6, 9-10
  equating of, 34-36
  format of, 10
  formula scoring, 4-5
  general characteristics of, 4-8
  item analysis, 39-41
  pretesting, 7, 10, 22
  purpose of, 4, 9
  quality control of, 7
  reliability of, 18-19, 39
  rescaling study, 38
  restructured, 10, 53-54
  restructuring, research on, 21-23
  review of, 5-6
  sample Aptitude Test, 7, 21
  scaled-score system, 32-34, 37-38
  specifications, research and statistical analysis in, 20-21
  speededness, 19-20
  statistical characteristics, 18-20, 39
    specifications, 20
  test analysis, 41-42
  validity of, 52-55, 57-60
Association of American Universities, 1
Association of Graduate Schools Committee on Testing, 1, 2
Attenuation, correction for, 44
Aukes, L. E., 61

Background questions, 27, 48-49, 53
Barrows, T. S., 58
Bass, A. R., 61, 98
Besco, R. O., 61
Biology Test, Advanced, 68-6[illegible]
  content, 68
  responses to background questions, 63-69
  validity data, 69
Biserial correlation, 7, 40-42, 45
Borg, W. R., 61
Bracht, G. H., 58
Braiding for scale stability, 37
Breland, H. M., 59
Brogden, H. E., 57

Campbell, D. T., 54
Capps, M. P., 61
Carlson, A. B., 20, 53
Carnegie Foundation for the Advancement of Teaching, 1
Chemistry Test, Advanced, 70-71
  content, 70
  responses to background questions, 70
  validity data, 71
Chen, C. C., 48
Clark, H., 61
Cleary, T. A., 59
Coaching studies, 20-21
Coffman, W. E., 13
College Entrance Examination Board, 2, 20
Comparability of scores, 9, 10, 32, 34, 36, 38-39
Computer Science Test, Advanced, 72
Construct validity, 52-54
Content validity, 52-53
Conway, Sister M. T., 61
Cooperative Graduate Testing Program, 1
Cornell Test of Critical Thinking, 22
Correlation
  of Advanced Test subscores, 30
  of Aptitude Test scores, 18-19, 30-31
  in test analysis report, 44-45
Council of Graduate Schools, 2
Creager, J. A., 56-57, 61, 69, 71, 77, 81, 88, 93, 99
Criterion-related validity, 52, 54-58
Criterion score for item analysis, 40-41
Cronbach, L. J., 55, 57, 58

Data interpretation questions, 14-15
Dawes, R. M., 61
Decosta, F. A., 61
Deductive reasoning, 66, 67
Delta scale, 40
  equating, 41
Descriptive statistics, 45-51
Discrete mathematics questions, 14-15
Dressel, P. L., 39, 43
Duff, F. L., 61

Eckhoff, C. M., 61, 75
Economics Test, Advanced, 73-74
  content, 73
  responses to background questions, 73-74
Education Test, Advanced, 74-76
  content, 74
  responses to background questions, 74-75
  validity data, 75-76
Efficiency of a test, 42, 44
Engineering Test, Advanced, 76-77
  content, 76-77
  responses to background questions, 77
  validity data, 77
Equating, 34-36, 41
  Advanced Tests, 35-36
  Aptitude Test, 35
  common item, 35
  delta, 41
  double-part-score, 37
  Levine equations, 35-36
  methods, 34-36
  Tucker equations, 35
Error of measurement, standard, 19, 29, 39, 43, 46
  Advanced Tests, 29
  Aptitude Test, 19
  error variance, 43
  raw score, 43
  scaled score, 43
Evaluation of evidence, 66
Evans, F. R., 21
Ewen, R. B., 61, 98

Factor analysis, 20, 53-54, 96
  Advanced Psychology Test, 96
  Aptitude Test, 20, 53
  construct validity, 53-54
  speededness, 20
Fishman, J. A., 56
Fiske, D. W., 54
Florida State University, 61
Ford, S. F., 59, 60
Formula scoring, 4-5
French Factor Kit, 22
French Test, Advanced, 42-47, 48
  content, 78
  responses to background questions, 78
  test analysis, 42-47

Geography Test, Advanced, 33, 36, 79
  content, 79
  responses to background questions, 79
  scaling, 33
  scaling of subscores, 36
Geology Test, Advanced, 80-81
  content, 80
  responses to background questions, 80
  validity data, 80-81
German Test, Advanced, 4, 25, 33, 82
  content, 82
  format, 25
  formula scoring, 4
  responses to background questions, 82
  scaling, 33
Gibbons, B. D., 61
Glass, G. V., 58
Goo, V., 61
Greb, D., 62, 76
Graduate Management Admission Test Program, 20, 22
Graduate Record Examinations Board, 2-3, 21, 60
  Research Committee, 21
Guessing, 4, 5, 21
  correction for, 4, 5
  research on, 21
  test instructions, 4

Hackman, J. R., 61, 98
Hansen, W. L., 61
Harlow, S. D., 62, 76
Harvey, P. R., 59, 61
High Level Math Usage Test, 10
Hilton, T. L., 59
History Test, Advanced, 83-84
  content, 83
  responses to background questions, 83-84
  validity data, 84
Holland, P. W., 51
Homogeneity, 26, 41, 43, 53
  item analysis criterion, 41
  reliability, effect on, 43
Houston, S. R., 62, 75
Humphreys, L., 59

Independent Student Testing Program, 1
Institutional Testing Program, 1
Intercorrelation, 41-43, 46, 53, 58
  in test analysis report, 41-43
Item analysis, 18, 20, 39-41
Item difficulty index (delta), 39, 40, 41
Item pool, 6, 40

Jackson, M., 61
Johnson, H., 84, 88
Jones, R. A., 61

Kendrick, S. A., 59
King, D. C., 61
Kuder, G. F., 39
Kuder-Richardson formula (20) for reliability, 39, 43

Lannholm, G. V., 56, 61, 84, 86, 91, 93, 97-98
Law, A., 61
Law School Admission Test Program, 20, 22
Letter sets, 64-66
Levine, R., 35-36
Linn, R. L., 59
Literature in English Test, Advanced, 85-86
  content, 85
  responses to background questions, 85-86
  validity data, 86
Logical diagrams, 17-18
Logical reasoning test, 10
Lorge, I., 61, 97

- *Madaus, G. F.*, 61
- *Marco, G. L.*, 61, 84, 86, 91, 93, 97-98
- Mathematics Test, Advanced, 87-88
  - content, 87
  - responses to background questions, 87-88
  - validity data, 88
- Mean score. *See* Summary statistics.
- Median, 42, 43, 47
- *Mehrabian, A.*, 61, 99
- *Messick, S.*, 58
- *Michael, W. B.*, 61
- *Michels, W. C.*, 92
- Multiple-choice format of test questions, 4, 10-17, 25-26
- Music Test, Advanced, 89-90
  - content, 89
  - responses to background questions, 89-90

- National Program for Graduate School Selection, 1
- National Science Foundation Graduate Fellowship Program, 37, 38
- *Newman, R. I.*, 61, 98
- *Nielsen, W.*, 62
- Normative data. *See* Descriptive statistics.

- *Office of Educational Research*, 61, 80
- *Olsen, M.*, 61

- Parallel forms of a test, 34
- *Pasanella, A. K.*, 58
- Percentile ranks, 46-48
- Philosophy Test, Advanced, 90-91
  - content, 90
  - responses to background questions, 90-91
  - validity data, 91
- Physics Test, Advanced, 92-93
  - content, [illegible]
  - responses to background questions, 92
  - validity data, 92-93
- *Pike, L. W.*, 21
- *Pitcher, B.*, 59
- Political Science Test, Advanced, 94-95
  - content, 94
  - responses to background questions, 94-95
- Population validity, 58-60
- Power test, 19, 29
- *Powers, D. E.*, 20, 53, 5[illegible]
- Predictability of graduate success, 56-58
- Predictive validity, 54-58
- Preliminary item analysis, 7, 40
- Pretesting, 7, 40. *See also* Aptitude Test.
- Profile Tests, 1
- Psychology Test, Advanced, 95-99
  - content, 95
  - research related to additional subscores, 95-96
  - responses to background questions, 96-97
  - validity data, 97-99
- *Pullias, C. M.*, 61

- Quality control, 7
  - item analysis, 40
  - test analysis, 41-45
- Quantitative ability measure, 9, 10, 14-16, 18-22, 55, 57, 58
  - changes in, 9

- Quantitative ability measure (continued)
  - content of, 14-16
  - content specifications, 16
  - descriptive statistics, 45-51
  - format, 10
  - reliability, 18-19
  - restructuring research, 21-22
  - types of questions, 14-16
    - data interpretation, 14-15
    - discrete mathematics, 14-15
    - quantitative comparisons, 15
- Quantitative comparisons, 15

- Reading comprehension sets, 9, 12-13
- Reliability, 18-20, 22, 28-30, 35-36, 39, 41, 43
  - Advanced Tests, 28-30
  - Aptitude Test, 18-20
  - effect on equating, 35-36
  - index of, 39
  - KR-20 formula, 39, 43
  - methods of determining, 39
  - relation to standard error of measurement, 43
- Rescaling study, 38-39
- *Richardson, M. W.*, 39
- *Roberts, P. T.*, 61, 69, 71, 84, 86, 88, 93, 101
- *Robertson, M.*, 62
- *Robinson, D. W.*, 62
- *Rock, D. A.*, 58, 62, 71, 88, 97
- *Roscoe, J. T.*, 62, 75
- *Rosenfeld, H.*, 62

- *Sacramento State College, Test Office*, 62
- Sample Aptitude Test, 7, 21
- Scaled-score system, 32-34, 36, 37-39
  - stability of, 37-39
- Scholastic Aptitude Test, 20, 21
- *Schrader, W. B.*, 61, 84, 86, 91, 93, 97-98
- *Schultz, M. K.*, 33
- Score scale, stability of, 36-37
- Sentence completion questions, 11-12
- *Shaffer, J.*, 62
- *Sistrunk, F.*, 62, 97
- Skewness, 42
- *Sleeper, M. L.*, 62
- Sociology Test, Advanced, 100-101
  - content, 100
  - responses to background questions, 100
  - validity data, 101
- Spanish Test, Advanced, 25, 101-102
  - content, 101
  - format, 25
  - responses to background questions 101-102
- Speededness, 26, 29, 43-44
- Spiraling of test forms for equating, 35
- Stability of the scale, 36-37
- Standard deviation. *See* Summary statistics.
- Standard error of measurement, 28, 39
- Standardization group, 1952, 32-33
- *Stanley, J. C.*, 59
- Statistical methods, 32-51
  - descriptive statistics, 45-51
  - equating, 34-36
  - error of measurement, standard, 39
  - item analysis, 39-41

- Statistical methods (continued)
  - reliability, 39
  - rescaling study, 38-39
  - scaled-score system, 32-34, 37-38
  - skewness, 42
  - speededness, 26, 43-44
  - spiraling (for equating), 35
  - stability of the scale, 36-37
  - subscore scaling, 36
  - test analysis, 41-46
- Subscores, 29-30, 36
- Summary statistics
  - Advanced Tests: subscores, 29, 43, 47
  - Advanced Tests: total scores, 28, 33, 37, 38, 42, 43
  - Aptitude Test, 37, 38, 48, 49, 50, 51, 58
- *Swineford, F.*, 39, 59
- *Swinton, S. S.*, 20, 53

- Test analysis, 7, 41-46
- Test assembly, 6
- Test development procedures, 5-6
- Test development staff, 5
- Test instructions, 4
- Testing standards, 7-8
- Tests of General Education, 1
- *Thompson, E.*, 84, 88
- *Tucker, L. R.*, 33, 35-36
- *Tully, G. E.*, 62

- Undergraduate Assessment Program, 2
- *University of Virginia, Office of Institutional Analysis*, 62

- Validity, 52-63
  - construct, 52-54
  - content, 52-53
  - criterion-related, 52, 54-58
  - definition of, 52
  - population, 58-60
  - predictive, 54-62
- Verbal ability measure, 10-14, 21, 55, 57, 58
  - content specifications, 13-14
  - discrete verbal questions
    - analogies, 11
    - antonyms, 11
    - sentence completion, 11-12
  - evolution of, 9
  - format of, 10
  - reading comprehension sets, 12-13
  - reliability, 18-19
- *Voorhees, H. R.*, 92

- *Wallace, A. D.*, 62
- *Wallmark, M. M.*, 38
- *Walsh, J. J.*, 61
- Watson-Glaser Test of Critical Thinking, 22
- *Wesman, A.*, 59
- *White, E. L.*, 62
- *White, G. W.*, 62
- *Wiggins, N.*, 61, 98
- *Wild, C. L.*, 59
- *Williams, J. D.*, 62, 76
- *Willingham, W. W.*, 52n, 58
- *Wilson, K. M.*, 55